Global epistasis and the emergence of function in microbial consortia

Diaz-Colunga, Juan; Skwara, Abigail; Vila, Jean C.C.; Bajic, Djordje; Sanchez, Alvaro

DOI
10.1016/j.cell.2024.04.016

Publication date
2024

Document Version
Final published version

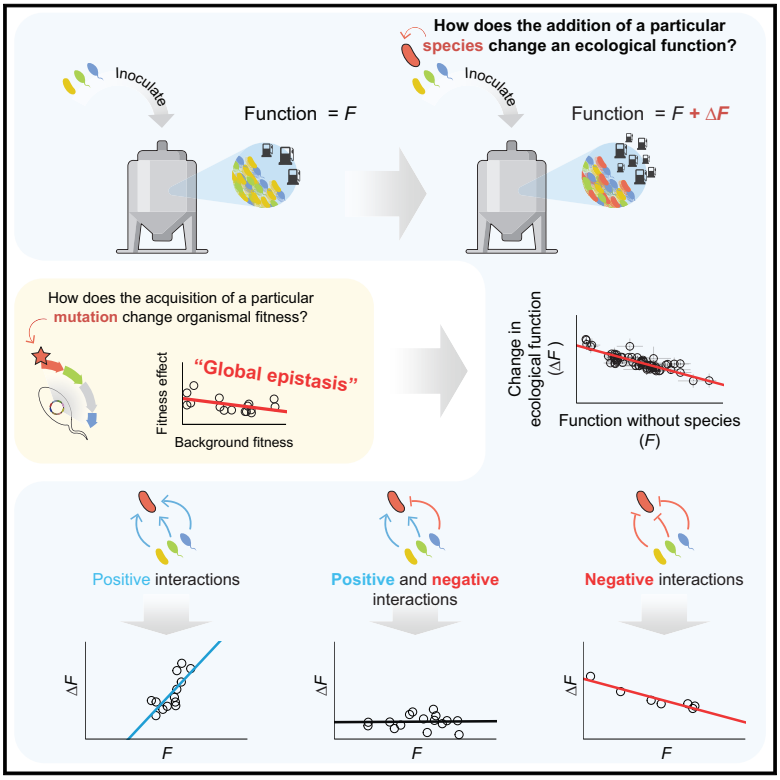
Published in
Cell

Citation (APA)
Diaz-Colunga, J., Skwara, A., Vila, J. C. C., Bajic, D., & Sanchez, A. (2024). Global epistasis and the emergence of function in microbial consortia. *Cell, 187*(12), 3108-3119.e30. https://doi.org/10.1016/j.cell.2024.04.016

Cell

Global epistasis and the emergence of function in microbial consortia

Graphical abstract



Highlights

- Simple statistical models predict the effect of a species on a community-level function
- These models mirror the patterns of global epistasis reported in genetics
- Ecological global epistasis emerges from widespread species-by-species interactions
- This phenomenon can be leveraged to optimize the functions of microbial consortia

Authors

Juan Diaz-Colunga, Abigail Skwara, Jean C.C. Vila, Djordje Bajic, Alvaro Sanchez

Correspondence

jdc@usal.es (J.D.-C.), d.bajic@tudelft.nl (D.B.), alvaro.sanchez@usal.es (A.S.)

In brief

Global epistasis, traditionally applied in genetics research, can be used to model complex microbial communities and predict the effects of a species on community-level functions.

Diaz-Colunga et al., 2024, Cell *187*, 3108–3119
June 6, 2024
https://doi.org/10.1016/j.cell.2024.04.016

CellPress

CellPress
OPEN ACCESS

Cell

Theory

Global epistasis and the emergence of function in microbial consortia

Juan Diaz-Colunga,[1,2,3,4,7,*] Abigail Skwara,[1,2,7] Jean C.C. Vila,[1,2,5] Djordje Bajic,[1,2,6,*] and Alvaro Sanchez[1,2,3,4,8,*]
[1]Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT 06511, USA
[2]Microbial Sciences Institute, Yale University, New Haven, CT 06511, USA
[3]Department of Microbial Biotechnology, National Center for Biotechnology CNB-CSIC, 28049 Madrid, Spain
[4]Institute of Functional Biology and Genomics IBFG-CSIC, University of Salamanca, 37007 Salamanca, Spain
[5]Department of Biology, Stanford University, Stanford, CA 94305, USA
[6]Department of Biotechnology, Delft University of Technology, Delft 2628 CD, the Netherlands
[7]These authors contributed equally
[8]Lead contact
*Correspondence: jdc@usal.es (J.D.-C.), d.bajic@tudelft.nl (D.B.), alvaro.sanchez@usal.es (A.S.)
https://doi.org/10.1016/j.cell.2024.04.016

SUMMARY

The many functions of microbial communities emerge from a complex web of interactions between organisms and their environment. This poses a significant obstacle to engineering microbial consortia, hindering our ability to harness the potential of microorganisms for biotechnological applications. In this study, we demonstrate that the collective effect of ecological interactions between microbes in a community can be captured by simple statistical models that predict how adding a new species to a community will affect its function. These predictive models mirror the patterns of *global epistasis* reported in genetics, and they can be quantitatively interpreted in terms of pairwise interactions between community members. Our results illuminate an unexplored path to quantitatively predicting the function of microbial consortia from their composition, paving the way to optimizing desirable community properties and bringing the tasks of predicting biological function at the genetic, organismal, and ecological scales under the same quantitative formalism.

INTRODUCTION

Microbial communities carry out critical functions in both natural and biotechnological settings, from nutrient cycling in the soils[1] to biofuel production in industrial biorefineries.[2] As hinted by the latter example, finding sustainable and economically viable alternatives to non-renewable resources relies on our ability to harness the useful capabilities of microbial communities. In biotechnological applications, even moderate increases in function can dictate the viability of a technology. Thus, it is often not enough to know whether a community will or will not carry out a given function; rather, we need to quantitatively and accurately identify optimal consortia. Given a list of candidate strains, which ones should we combine together if we wish to maximize the amount of biofuel produced in a bioreactor,[2] the elimination of toxic compounds,[3,4] or the suppression of a pathogen?[5] To adequately answer these questions, we must be able to quantitatively and accurately predict how including each of those strains in a consortium will affect the desired function (Figure 1A).

The primary drawback of current approaches for addressing this problem is that they struggle with interactions. The contribution of a species to any desired community function emerges from a complex web of interactions with the other community members, such as competition for resources, metabolic cross-feeding, or ecological effects on gene expression, among others.[7] This contribution is thus often contingent on which other species are present[7] (Figure 1A). The number of such potential interactions grows exponentially with community size, making them highly challenging to characterize in practice. For this reason, past models have only been successful at identifying optimal consortia for very small communities where interactions are sparse,[8–14] following statistical approaches that may encounter scalability issues in more complex situations. For larger communities, predictive ecological frameworks have been primarily concerned with explaining broad-scale ecological patterns, such as those that exist between biodiversity and ecosystem functioning (focusing generally on its productivity).[15–21] These models have provided valuable insights into the functioning of natural ecosystems. However, they lack the fine-grained predictive power required to identify which particular combination of microbial species will optimize any desired community function (e.g., the production of a target molecule[2,13]). A general, scalable solution to this latter question requires a different approach to handling interactions.[9,11–14]

Interactions between biological components complicate the construction of predictive models in other areas of biology besides ecology. In genetics, for instance, the contribution of a

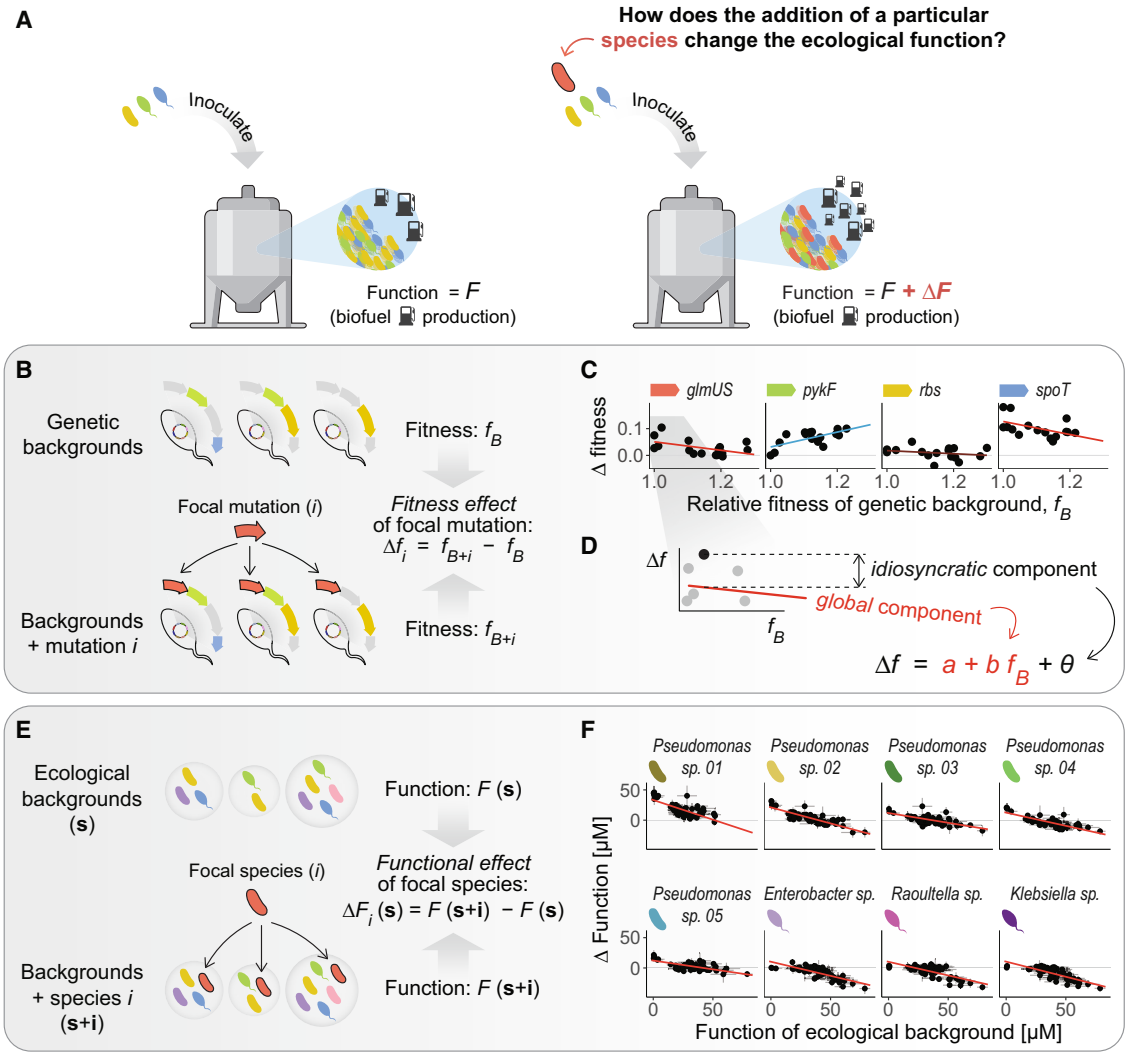


Figure 1. An ecological parallel to global epistasis

(A) The central challenge of this work is predicting how a collective ecological function (in this illustration, the production of bioethanol) will change with the addition of a particular species.

(B and C) Recent work in quantitative genetics has found that the fitness effect of a mutation is often predicted by the fitness of the genetic background where it arises, following simple regressions. This phenomenon has been termed *global epistasis*. The slope and intercept of the regression vary across different mutations. Data from Khan et al.[6]

(D) The fitness effect of a mutation can thus be broken down into two components: first, a *global* component that is predictable from the regression, and second, an *idiosyncratic* component that is not predictable from the background fitness, represented by the residuals of the fits.

(E) We hypothesized the existence of an ecological parallel to global epistasis, where the addition of a focal species to a community might induce a change in a community-level function that can be predicted from the function of the background community where the focal species is added.

(F) We inoculated different combinations of eight bacterial isolates in synthetic liquid medium. The functions of these consortia were quantified as the level of pyoverdines secretion after 48 h of incubation (STAR Methods, Figure S1). We found that the "functional effect" of adding a species to an "ecological background" was well predicted by a simple regression linking it to the function of the background community. The background functions and functional effects were quantified independently in three biological replicates; here we represent their averages (dots) and standard deviations (error bars).

See Figures S1–S3, S5–S9, and S22.

locus to a quantitative phenotype depends on its genetic background via genetic ("epistatic") interactions with other loci.[22] Epistatic interactions can be widespread, and they may also be highly complex, including pairwise as well as higher-order effects.[23–28] Encouragingly, recent research has shown that the aggregate effect of these genetic interactions often leads to the emergence of simple statistical patterns, where the fitness effect of a mutation is well predicted by the fitness of its genetic background (Figures 1B–1D). The emergence of these simple statistical patterns is a manifestation of "global epistasis"[6,29–40]—a phenomenon that includes (but is not limited to) the common observation that beneficial mutations have

smaller fitness effects in higher-fitness genetic backgrounds (“diminishing returns epistasis”). As an illustration of global epistasis, in Figure 1C, we reproduce previous results showing that the fitness effects of four *Escherichia coli* mutations are well predicted by the fitness of their genetic backgrounds through simple linear regressions.[6] These global epistasis patterns can be determined from just a small number of empirical observations, and theory has shown that they can be mechanistically interpreted in terms of gene-by-gene (g×g) interactions.[38,39] Recent studies have leveraged global epistasis to develop highly promising methodologies for inferring full genotype-phenotype maps in large combinatorial spaces from just a subset of measurements.[41–45]

Inspired by recently established parallels between genetic and functional ecological interactions,[46–49] here we hypothesize that an ecological analog to global epistasis might exist, where the aggregate effect of all pairwise and higher-order ecological interactions between species may be captured by simple regression models that predict how including a species in a community will affect its community-level function (Figure 1E). If such simple, predictive models existed and could be interpreted in terms of interspecies interactions (as is the case for global epistasis in genetics), this would unlock our ability to predictively connect species-level composition to quantitative function across a wide variety of ecological systems.

To evaluate the merits of this hypothesis, we conducted a series of experiments using synthetic microbial communities and also examined previously published data of bacterial, algal, and even plant ecosystems, under distinct environmental conditions and for a variety of collective functions. We found that a parallel concept to global epistasis can indeed be formulated in all these cases. Here, we show that this allows us to build predictive models to optimize ecological function that are simple, general, and highly interpretable. Furthermore, we extend previous theoretical results from the field of genetics to demonstrate that ecological global epistasis-like patterns can be quantitatively linked to species-by-species (s×s) interactions. Our findings argue that the same general formalism can be applied to predict biological function across widely different scales of organization, from molecules and organisms to ecological communities.

RESULTS

An ecological parallel to global epistasis predicts the functional effect of a species

Based on our hypothesis, we could define the functional effect of a species by analogy with the fitness effect of a mutation, i.e., as the difference in function between two communities that differ solely in the presence or absence of such species (Figure 1E). If our hypothesis were correct, we should observe that the functional effect of a focal species would be well predicted by the function of the (background) community where we include it, following simple statistical models similar to those observed in genetic global epistasis.

To test our hypothesis, we built a small library of eight soil bacterial isolates (STAR Methods), from which one could potentially assemble 255 different consortia based on the presence/absence of each member. The community-level function we studied was the net production of pyoverdines. This represents a good test case for an ecological function that is sensitive to species interactions, as pyoverdine secretion is known to respond to intra-species signaling[50] and is also often controlled by population size via quorum sensing.[51] Five of our species were *Pseudomonas* strains that produce pyoverdines in monoculture, while the remaining three were non-producing *Enterobacteriaceae* (Figure S1).

From this library, we assembled a subset ($N = 164$) of all possible unique species combinations, including an approximately even representation of pairs, trios, etc. (Methods S1). Each of these consortia can be represented by a vector $\mathbf{s}$, which encodes the presence/absence of each species i ($s_i = 1,0$). To form each consortium, we inoculated every member at a fixed inoculum density in minimal liquid growth medium (STAR Methods, Figure S1). We then incubated our consortia for 48 h and measured their function ($F(\mathbf{s})$) as the concentration of pyoverdines in the spent media (STAR Methods, Figure S1). The assembled consortia exhibited high variation in functional levels, with pyoverdine concentrations ranging from 0 to 70 μM (Figure S1). In Figure 1F, we plot the functional effect of each species ($\Delta F_i(\mathbf{s}) = F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s})$ for species i, Figure 1E) against the function of its background consortia ($F(\mathbf{s})$).

Consistent with our hypothesis, we found that the functional effects of all species in different community contexts were well predicted by simple relationships of the form $\Delta F_i(\mathbf{s}) = a_i + b_i F(\mathbf{s}) + \theta_i(\mathbf{s})$ (Figure 1F). We call this expression the “functional effect equation” (FEE) of species i. The intercepts (a_i) and slopes (b_i) of the FEEs differ across species, suggesting that they are determined by specific interactions between each individual species and the rest of its ecological partners (Figure S2). The terms $\theta_i(\mathbf{s})$ (i.e., the residuals of the fits) capture the component of said interactions that is not predictable from the statistical model itself. Consistent with our initial hypothesis, these results indicate that a parallel to global epistasis exists in this ecological system, which predicts how including a particular species into different community contexts will affect their community-level function.

The ecological parallel to global epistasis is ubiquitous across a wide range of ecological communities

To determine how general this phenomenon might be beyond our particular experimental setting, we re-analyzed a collection of already published datasets where the quantitative relationship between community composition and function had been measured in a similar manner. Table S1 summarizes the datasets we considered, all of which include multiple combinatorial assemblages of species from candidate pools of 4 to 25 species. These datasets include synthetic bacterial communities formed by either Gram-negative or Gram-positive bacteria,[13,47,52] but also plant[53] and phytoplankton[54] communities. These sets of communities were assembled under widely different ecological conditions, including the number of organismal generations, the type and frequency of resource addition, and the form of propagation. The functions were also different in each case, ranging from the production of biomass to the net metabolic activity and from the secretion of specific enzymes to the degradation of environmental polymers. As shown in Figure 2, simple

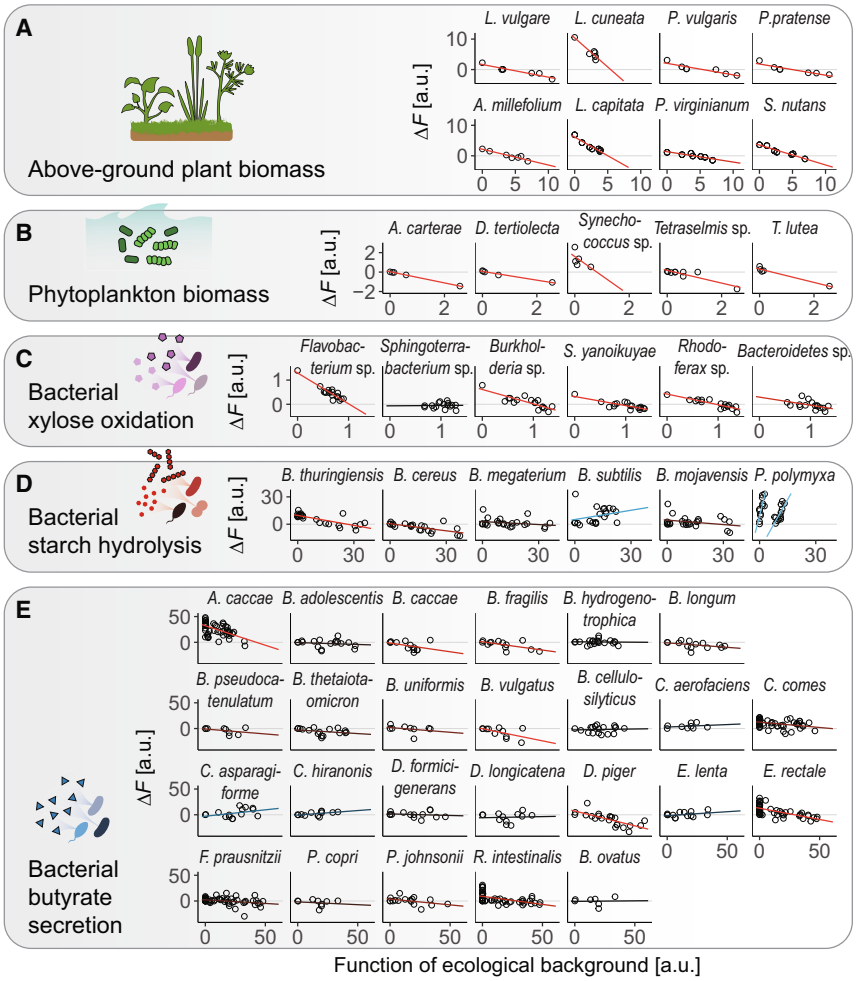


Figure 2. Ecological global epistasis across different communities and functions

The functional effect of a focal species can often be predicted from a simple regression linking it to the function of the background community where it is included (red lines: negative slopes, blue lines: positive slopes, black lines: near-zero slopes). This is observed across communities formed by very different organism types, under different ecological conditions, and for different collective functions (Table S1).

(A) Data from Kuebbing et al.[53]

(B) Data from Ghedini et al.[54]

(C) Data from Langenheder et al.[52]

(D) Data from Sanchez-Gorostiaga et al.[47]

(E) Data from Clark et al.[13]

See Figures S3–S9 and S22.

statistical models predict the functional effect of species across all datasets we investigated, explaining on average ~75% of the variance in a species' functional effect (Figure S3).

Most species (~50% across all datasets) display negatively sloped FEEs (red lines in Figure 2). This trend is also commonly observed in population genetics: the fitness effect of a mutation most often becomes either less beneficial ("diminishing returns") or more deleterious ("increasing costs") as the fitness of the genetic background increases.[6,29,30,33,35,36] Often, species increase community function when they are included in low-performing ecological backgrounds but decrease it when the background function is high. About 45% of all species exhibit near-zero slopes (black lines in Figure 2). Note, however, that these patterns are also informative for predictive purposes. The magnitude of the deviations from the FEE (even if its slope is zero) is useful to discern between (1) species whose contribution to ecosystem function is additive and largely independent of their ecological background (i.e., those species for which the residuals are small) or (2) species whose contribution to the community function depends on the specific composition of their ecological background in a very idiosyncratic manner (i.e., those with large residuals). Finally, a smaller number of species (~5%) exhibit positively sloped FEEs (blue lines in Figure 2), becoming more beneficial (or less deleterious) in backgrounds with higher functions. We refer to these patterns as "increasing returns" (or "decreasing costs").

The reader will have noticed that *P. polymyxa* displays two distinct types of functional effects on the function (amylolytic rate) of its background consortia, each described by a different FEE (Figure 2D, rightmost panel). Closer examination of this species indicates that these two "branches" are defined by the presence or absence of a second species (*B. thuringiensis*) in the ecological background (Figure S4). We will address this irregularity in more detail later in the text.

In Figures 1F and 2, we have represented the functional effect of a species versus the function of the ecological background. It is important to note that this type of representation can lead to biased estimates of the strength of global epistasis. In short, this is because the variable represented on the y axis (ΔF) explicitly depends on the variable represented on the x axis ($F(\mathbf{s})$), and correlations between them can result from co-varying measurement errors.[37,55] Recent work in genetics has addressed this issue by directly regressing the background fitness against the fitness of the genotype where a focal mutation has been added via total least-squares regression.[37] In Figure S6, we adopt the same strategy, directly representing $F(\mathbf{s}+\mathbf{i})$ against $F(\mathbf{s})$, and we find slopes different from 1 for a majority of species (Figure S7). Further analysis indicates that most of the empirical FEEs we find across datasets are not a simple consequence of measurement noise (Methods S1; Figures S7–S9).

Emergence of global functional interactions from pairwise interactions between species

Our analyses suggest that ecological global epistasis is ubiquitous across a wide range of communities. How should we explain this mechanistically? And what factors determine

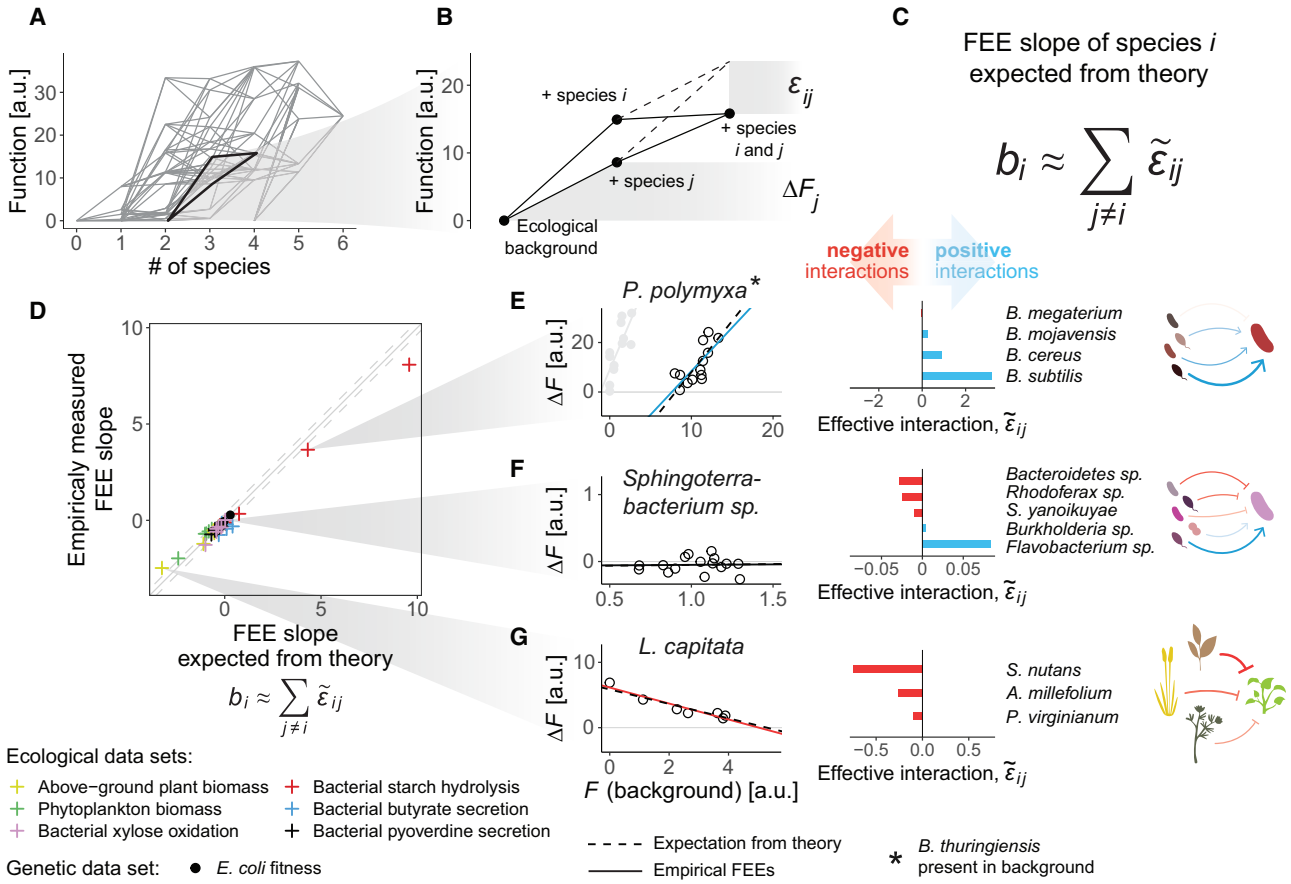


Figure 3. Effective pairwise ecological interactions explain FEE slopes

(A) Mapping between the number of species and the functions of the communities in the Sanchez-Gorostiaga et al. experiment.[47] Each node represents a consortium, and edges connect consortia that differ in the presence of a single particular species.

(B) Detail showing an example where the inclusion of two species (*i* and *j*) in an ecological background results in lower function (solid lines) than the additive expectation (dashed lines). This difference (ϵ_{ij}) indicates a functional interaction between the two species (either a direct one or an indirect one, e.g., mediated by other community members). ΔF_j is the functional effect of species *j* on the ecological background.

(C) Based on theoretical results from quantitative genetics, we hypothesized that the FEE slope of a species *i* may be explained by a sum of its effective interactions with every other species *j*. The effective interaction of species *i* with species *j* (denoted $\tilde{\epsilon}_{ij}$) is defined as $\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \langle \Delta F_j \rangle / \sum_{k \neq i} \langle \Delta F_k \rangle^2$ (see Box 1), where the averages are taken across all possible ecological backgrounds where both species *i* and *j* are not present.[38,39]

(D) We quantified all effective interactions for every species and dataset (STAR Methods) and used them to estimate the expected FEE slopes. The estimated slopes are in agreement with the empirical fits in Figure 2 ($R^2 = 0.98$).

(E–G) We show three examples of species with positive, zero, and negative slopes. Slopes are explained by the sign and magnitude of the effective interactions between the focal species and its ecological partners.

See Figures S10–S16.

whether a species will exhibit diminishing or accelerating returns? In quantitative genetics, global epistasis patterns may emerge from pairwise and higher-order epistatic interactions, with the sign and magnitude of such interactions dictating the strength and shape of global epistasis.[38,39] If a similar relationship held in ecology, we could connect the FEE of a species to its functional interactions with other community members. In our own experiment of pyoverdine-secreting communities, as well as in the other datasets we analyzed, we found that species' functional effects (Figure S10) and functional interactions between species (Figure S11) were of variable sign and magnitude. We thus hypothesized that the different FEEs we found across species could be explained by this variation.

To test this hypothesis, we define the "effective interaction" between species *i* and *j* as $\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \langle \Delta F_j \rangle / \sum_{k \neq i} \langle \Delta F_k \rangle^2$ (see Box 1). Here, $\langle \epsilon_{ij} \rangle$ denotes the average functional interaction between species *i* and *j*, that is, the average difference in function between a consortium where both species are present with respect to the additive expectation from each species' individual functional effect (Figure 3B). In turn, $\langle \Delta F_j \rangle$ denotes the average functional effect of species *j* across all ecological backgrounds that do not contain species *i* nor *j* (Figure 3B). This definition follows from recent work in quantitative genetics (we refer readers to Reddy and Desai[38] and Diaz-Colunga et al.[39] for a detailed derivation of this expression, here summarized in Methods S1), which found that the slope b_i of the global epistasis regression for a

Box 1. Glossary

Global epistasis. In genetics, the fitness effect of a mutation has often been found to be correlated with the fitness of the genotype where the mutation arises (usually called the "genetic background"). This phenomenon has been termed "global epistasis".[6,29–40]

Ecological background. An ecological community to which a focal species *i* can be added. This is analogous to how a genetic background is usually defined, as a genotype that does not carry a given focal mutation.

Functional effect. We define the "functional effect" of a focal species as the quantitative difference in ecological function between two consortia that differ solely in the presence or absence of that focal species. Note that this definition is analogous to that of a mutation's "fitness effect" (or "phenotypic effect") in the context of genetics. We denote the functional effect of species *i* as ΔF_i.

Functional interaction. We consider that two species engage in a "functional interaction" when including the two of them in a consortium produces a change in a quantitative ecological function that deviates from the sum of their two separate functional effects. If this difference is positive (negative), we say the two species engage in a positive (negative) functional interaction. We denote the functional interaction between species *i* and *j* as ϵ_{ij}.

Effective interaction. By the above definition, the functional interaction between two species may vary depending on which other species are present or absent in the consortium. We mathematically define an "effective interaction" between two species *i* and *j*, denoted $\tilde{\epsilon}_{ij}$ as

$$\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \frac{\langle \Delta F_j \rangle}{\sum_{k \neq i} \langle \Delta F_k \rangle^2} \quad \text{(Equation 1)}$$

where $\langle \epsilon_{ij} \rangle$ represents the functional interaction between species *i* and *j* averaged across all possible ecological backgrounds, and $\langle \Delta F_j \rangle$ represents the average functional effect of species *j* across that same set of ecological backgrounds.[39]

mutation *i* could be approximated as the sum of its effective interactions with all other mutations *j*. Applying the theory of global epistasis to community function, we reasoned that the slope (b_i) of the FEE for species *i* may thus be written as (Figures 3A–3C):

$$b_i \approx \sum_{j \neq i} \tilde{\epsilon}_{ij} \quad \text{(Equation 2)}$$

From the expression above, we notice that species that engage primarily in negative effective interactions will tend to exhibit diminishing returns, whereas those that engage in positive effective interactions will exhibit increasing returns.[39] Note that, as defined, negative and positive effective interactions refer to the average impact of a species pair on a community-level function across all ecological backgrounds and need not correspond to a direct positive or negative ecological interaction between the two species.

To test whether this equation could explain the FEEs in Figures 1 and 2, we estimated the effective interactions between all pairs of species in those experiments (STAR Methods). We found that Equation 2 does an excellent job at explaining FEE slopes across datasets (Figures 3D and S12), and it helps us identify the mechanistic basis behind the different patterns of global epistasis for each species. For instance, *P. polymyxa* exhibits increasing returns, and this pattern is explained by its predominantly positive effective interactions with other community members (Figure 3E). In turn, the flat slope exhibited by *Sphingoterrabacterium* sp. (Figure 3F) arises from its evenly balanced positive and negative effective interactions, while the negative slope of *L. capitata* (Figure 3G) can be attributed to its negative effective interactions with all other species. Beyond these three examples, in Figure S13, we show the sign and magnitude of all effective interactions between every pair of species across all datasets we examined.

Equation 2 also allows us to rationalize the two branches observed in the FEE for *P. polymyxa*. As shown in Figures 2D and S4, this species exhibits two distinct types of functional effects. The apparent split of this FEE into two branches can be explained by the particular interaction structure of *P. polymyxa* with other community members. Specifically, *P. polymyxa* exhibits a strong negative interaction with a second species (*B. thuringiensis*) and positive interactions ($\tilde{\epsilon} > 0$) with all others. The presence or absence of this second species in the ecological background determines the two branches observed in this FEE. This phenomenon can be analyzed in light of Equation 2 and the complementary equation that explains the intercept (Figure S14; Methods S1).

The definition of an effective interaction, $\tilde{\epsilon}_{ij}$, results from averaging the interaction between a pair of species *i* and *j* across all ecological backgrounds where they may be included. In general, the interaction between species *i* and *j* may vary depending on the presence or absence of additional community members, which they may engage in higher-order interactions (HOIs) with. Figure 3 shows that averaging this variation is generally sufficient to provide good estimates for the empirical FEE slopes, but this does not mean that HOIs do not exist in the datasets we considered. If HOIs were truly absent, we should be able to accurately estimate FEE slopes by analyzing just the one- and two-species communities, but this is not the case in general (Methods S1; Figures S15 and S16).

The analysis presented in Figure 3 suggests that the FEE of a species may potentially be explainable mechanistically in terms of its average ecological interactions, which in turn can be understood at the molecular level. For instance, as discussed above, the positive slope of *P. polymyxa* results from several positive effective interactions with its ecological partners. These positive interactions have a known molecular basis: *P. polymyxa* is a biotin auxotroph whose growth is facilitated via cross-feeding by other members of the consortia.[47] This observation highlights the utility of defining effective functional interactions between species in order to bridge the gap between molecular-level mechanisms and the emergence of community-level functions.

FEEs can be leveraged to predict community-level functions

At the outset of this paper, we argued that the ability to predict the functional effects of a set of candidate species would allow us to quantitatively link the composition and function of any consortium one may form with them. In fact, FEEs make it possible to identify which of those consortia will optimize a community function. A simple visual inspection of the FEEs can be useful for this

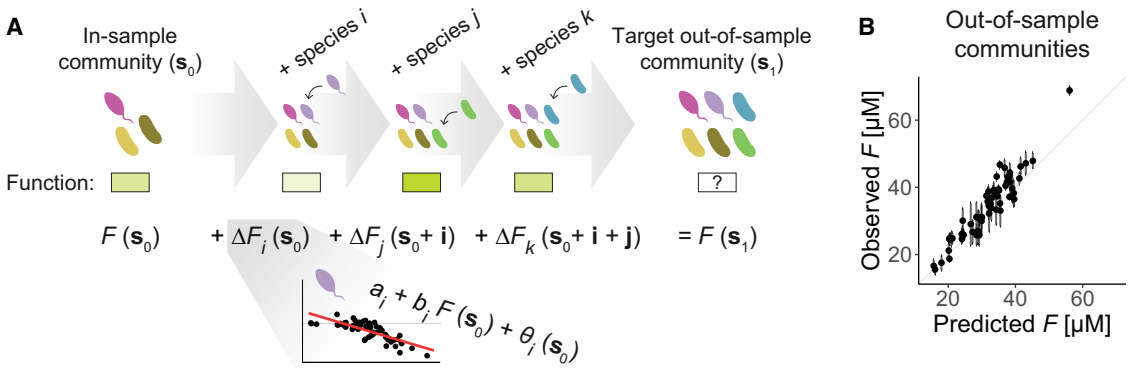


Figure 4. Ecological global epistasis can be leveraged to predict community functions

(A) We hypothesized that iteratively concatenating the functional effects of a set of species (here species *i*, *j*, and *k*) could serve to predict the function of an out-of-sample community ($\mathbf{s}_1$) from that of an in-sample community ($\mathbf{s}_0$) (STAR Methods).

(B) We evaluated the performance of this method by assembling 61 new consortia (which served as the out-of-sample test set of communities) and comparing their predicted and measured levels of pyoverdines secretion. We found a good agreement (R^2 = 0.80) between the observations and the predictions. Dots and error bars represent means and standard deviations across two biological replicates.

See Figures S17–S21.

purpose. For instance, species that exhibit patterns of increasing costs may be discarded due to their detrimental effect on community function, which is more pronounced in higher-performing backgrounds. On the other hand, species exhibiting accelerating returns could be more promising: they would act as functional "boosters" by bringing up the function of their background community, even more strongly when the background function is already high. Because FEEs are easily and intuitively interpretable, this simple observation can help us narrow down the list of potentially desirable species. Beyond these straightforward guidelines, we reasoned that FEEs could serve to obtain quantitative predictions of community function based on composition and thus to identify optimal consortia.

Perhaps the simplest approach to predict the function of a consortium would be to iteratively concatenate the FEEs of all its members: if we wanted to predict the function of a particular consortium (denoted $\mathbf{s}_1$, Figure 4A), we could start by identifying a consortium $\mathbf{s}_0$ whose function had been measured empirically, such that all species in $\mathbf{s}_0$ were also present in the target consortium $\mathbf{s}_1$. In the example in Figure 4A, the target consortium $\mathbf{s}_1$ contains all species in $\mathbf{s}_0$, plus three additional species, which we generically denote *i*, *j*, and *k*. We could first use the FEE for species *i* to estimate the functional effect of including species *i* in the starting consortium $\mathbf{s}_0$, simply as $\Delta F_i(\mathbf{s}_0) \approx a_i + b_i F(\mathbf{s}_0)$. Then, the function of the resulting consortium ($\mathbf{s}_0 + \mathbf{i}$) would be estimated as $F(\mathbf{s}_0 + \mathbf{i}) = F(\mathbf{s}_0) + \Delta F_i(\mathbf{s}_0)$. We can next estimate the functional effect of including species *j* in this consortium from the FEE of species *j*, as $\Delta F_j(\mathbf{s}_0 + \mathbf{i}) \approx a_j + b_j F(\mathbf{s}_0 + \mathbf{i})$, and iterate this process one more time to estimate the effect of including species *k*: $\Delta F_k(\mathbf{s}_0 + \mathbf{i} + \mathbf{j}) \approx a_k + b_k F(\mathbf{s}_0 + \mathbf{i} + \mathbf{j})$.This will ultimately give a prediction for the function of the target consortium $\mathbf{s}_1$ (Figure 4A).

It is important to note that this procedure will yield quantitatively different predictions depending on the order in which FEEs are concatenated (e.g., $i \to j \to k$, $j \to k \to i$, etc.). One option is to consider every possible order of concatenation and average the predictions across them. As an alternative, we propose a method to infer the residuals of the FEEs, such that concatenating FEEs of the form $\Delta F_i(\mathbf{s}) \approx a_i + b_i F(\mathbf{s}) + \theta_i(\mathbf{s})$ (with $\theta_i(\mathbf{s})$ representing the residual of the FEE for species *i* in background $\mathbf{s}$) results in identical predicted values regardless of the order of concatenation—details on this residual inference procedure can be found in Methods S1.

Despite its conceptual simplicity, this approach yields excellent results at predicting the function of newly assembled communities. For instance, in Figure 4B, we generated predictions for the community-level pyoverdine production of a set of 61 new consortia, none of which had been assembled in our previous experiment (i.e., in Figure 1F). We then went back to the laboratory, assembled those consortia *de novo*, and measured their empirical function (STAR Methods, Figure S1). The agreement between our predictions and the empirical measurements was excellent (R^2 = 0.80, Figure 4B). The success of this simple approach is also remarkable across all of the other datasets we had previously analyzed (Figure S17). Inferring FEE residuals as explained in Methods S1 is not strictly necessary for prediction (averaging across all possible orders of FEE concatenation is enough for out-of-sample prediction), but it systematically resulted in increased prediction accuracy across all datasets (Methods S1). The method was also able to successfully identify optimal consortia in our pyoverdines experiment as well as in all the other datasets, even when FEEs were fit to very few data points (Figures S18 and S19). Our method for predicting community functions yielded more precise predictions than alternative state-of-the-art statistical models for the same purpose,[14] particularly when predicting the function of the highest- and lowest-performing consortia (Figure S20; Methods S1).

To further examine the performance of this predictive method, we simulated a set of mappings between community composition and function. For each simulation, we considered ecological interactions of variable sign, strength, and order. We found that the sign and magnitude of the underlying pairwise interactions

between species had a minor effect on the performance of the method, but the quality of the predictions declined in the limit where high-order interactions were both very strong and widespread (Figure S21; Methods S1).

The observation that this very simple statistical approach can quantitatively predict community-level functions highlights the utility of characterizing the FEEs, as they provide a simple and interpretable pathway toward linking community composition and function. This also suggests that more sophisticated statistical machinery, recently developed in genetics to quantitatively predict phenotypes from genotypes,[41–45] could be extended to ecology to predictively connect the composition and function of species assemblages.[13,48]

DISCUSSION

There is currently a growing interest in engineering microbial communities to carry out biotechnologically relevant functions, such as the fermentation of food and drinks,[56] the production of biofuel,[2] the degradation of environmental pollutants,[4] or the exclusion of pathogens.[5] Yet, building predictive models of community function that integrate the full complexity of functional and ecological interactions is extremely challenging. Such models have only been built in a small number of case studies,[8–10,13] but their parameterization required exhaustive empirical work that was highly specific to the species, environmental conditions, and functions being studied. Machine learning strategies are more scalable,[11,12] but extracting interpretable biological information from them is generally difficult. As an alternative, a common coarse-grained description of ecological communities involves reducing community structure to a scalar metric of biodiversity.[21,57] When averaged across multiple communities, biodiversity is often correlated with ecosystem function, but the variation is typically high, and the specific form of this relationship can vary in different ecological contexts. The relationships between biodiversity and ecosystem function provide valuable insights in natural settings, but they cannot resolve which specific consortia one must form to maximize a function of interest.

Our findings suggest that these limitations can be overcome by leveraging an ecological analog to the concept of global epistasis, originally formulated in the context of genetics. We have shown that simple linear regression models FEEs predict the functional effect of a species in a given background community, and they emerge ubiquitously across very different ecological conditions and collective functions. Much like genetic global epistasis patterns, FEEs can be characterized from a small subset of empirical observations (even in large combinatorial spaces) without the need for detailed information on the molecular mechanisms governing the interaction between every pair of species.

We propose that these FEEs may be interpreted as representations of emergent, coarse-grained species-by-community (s×C) interactions. Historically, the study of ecological interactions has broken them down as the sum of pairwise interactions (s×s) and HOIs (e.g., s×s×s, etc.).[46,49,58–62] This logic has paralleled the way in which genetic interactions have been traditionally partitioned, as the sum of pairwise gene-by-gene (g×g) interactions, third order interactions (g×g×g), fourth order, and so on.[23–28] The observation of global epistasis in genetic systems has revealed that epistasis can be instead partitioned into a global component, described by a linear regression between the fitness effect of a mutation and the fitness of the genetic background, and an idiosyncratic component described by the residuals of this fit. Building on recent analogies between genetic and functional ecological interactions,[46,47,49] here we have demonstrated that the latter can be partitioned in the same manner, as the sum of a global s×C interaction described by the FEEs, and an idiosyncratic component captured by the residuals. Furthermore, we have shown that the emergence of these global s×C interactions can be explained in terms of specific species-by-species interactions, expanding on recent theoretical results from the field of quantitative genetics.[38,39]

Finally, we have shown that even a very simple statistical approach, based on using FEEs to predict community function out of sample, can predict ecological function with high precision. This observation paves the way for the development of more advanced methods that leverage FEEs to predict and optimize community-level functions (e.g., based on current methodologies that aim to predict organism-level phenotypes from genotypes[41–45]).

Limitations of the study

Below, we highlight some of the limitations of our results, together with potential avenues for further investigation.

Temporal variation of community composition and function

The intrinsic ecological dynamics of a community can be generally expected to change its composition and function over time, including the possibility of one or more species being competitively excluded from a consortium[63] (but note that even transient species that eventually go extinct can induce long-lasting effects on community-level properties[64]). The magnitude and sign of species interactions may also be sensitive to temporal variation.[65] Thus, the FEE for any given species may be expected to change depending on the time point chosen to harvest the function of interest. Perhaps a reasonable expectation is that species that are often competitively excluded will tend to exhibit near-zero functional effects across backgrounds if the function is harvested many generations after inoculation. Both in our own experiments and in the datasets we re-analyzed, community functions were quantified after a single batch incubation. This type of scenario is common in biotechnological applications: food fermentation or biofuel production typically occurs in closed bioreactors over a predetermined time period, much like crop fields are generally harvested once plants reach maturity. In other cases, one might want to engineer ecological communities to be propagated in time, subject to environmental variation or the influx of invader species. It is unclear whether adding a new species to such communities (as opposed to simultaneously co-inoculating the new species together with the other members of its ecological background) may affect their ecological functions in a predictable manner.

Effect of inoculum size

In our analyses, we have assessed the effects of including a given species in a community at a fixed density. It is reasonable

to expect that the same species, inoculated at different starting population sizes, will have different effects on community function. Perhaps a reasonable expectation is that a particular species, when inoculated at very low initial density, will have a minor effect on the function of the consortium when grown over a single batch period, leading to a FEE with a slope and intercept of zero. By contrast, if the same species is inoculated at an initial density that is much higher than those of the other members of the consortium, the consortium may approximate the behavior of a monoculture of the new species, which would then exhibit a FEE slope close to -1. Additional theoretical and empirical work will be needed to assess how inoculum size may affect the FEE of a given species.

Complexity of the target ecological function

Here we have studied qualitatively simple community functions, which can be carried out to some degree by individual community members in isolation. Many applications may require the optimization of more complex functions (e.g., the simultaneous secretion of two or more metabolites or the removal of multiple environmental pollutants).

Effect of interspecies interactions

Our work indicates that the sign and magnitude of the interactions between community members do not affect the predictive power of FEEs, except when high-order interactions are strong and widespread (Figure S21; Methods S1). Recent work has suggested that high-order interactions may be relatively sparse in microbial communities,[14,66] but the generality of this observation remains an open question.

Effect of environmental variation

Interactions between species are known to be affected by environmental abiotic factors, such as temperature,[67] pH,[68] or nutrient availability.[69] Thus, we may expect the FEE of a species to be sensitive to these and other environmental variables, much like the global epistasis pattern of a mutation may change when environmental conditions are altered.[37,70,71] Further work is needed to assess how species-by-environment interactions may shape FEEs and our ability to predict ecological function across environments.

Scalability of FEE-based prediction approaches

In this study, we have analyzed ecological communities made up by relatively few (4–25) member species. Empirical datasets where a quantitative function is mapped to the presence or absence of a set of species are still scarce, possibly because the bottom-up assembly of large collections of microbial consortia remains a labor-intensive process. Future work will need to explore whether FEEs may be characterized and ultimately leveraged to predict ecological function in larger combinatorial spaces.

Alternative statistical methods for predicting ecological function

To illustrate that FEEs could be leveraged to predictively connect the composition and function of microbial consortia, we deployed the simplest approach we could think of, which relies on concatenating the FEEs of all member species (Figure 4). There are many ways to improve upon this methodology, in particular related to self-consistency (i.e., dealing with the lack of commutativity in the order in which the FEEs are concatenated). The strategy we have followed here is merely a first pass, but we hope that its success at predicting community functions out of sample (already surpassing state-of-the-art approaches in community ecology[14]) will stimulate the development of more sophisticated statistical methodologies leveraging global epistasis for learning ecological structure-function maps. Several such strategies have been developed in genetics for inferring genotype-phenotype maps from subsets of empirical observations.[41–45]

Clearly, the limitations of our study highlight that additional theoretical and empirical work will be needed to exploit the full potential of global epistasis for engineering microbial consortia. This is likely to require generating combinatorial datasets mapping community structure to function with larger sets of species than the ones currently available. Yet, our observation that global epistasis-like patterns emerge in ecological communities has clear and immediate practical uses for community-level engineering. Perhaps most importantly, our results argue that predictively linking biological structure to function can be attained through a common, general quantitative formalism across scales of organization—from molecules and organisms to ecological communities.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Bacterial species isolation from environmental samples
- METHOD DETAILS
  - Combinatorial community assembly
  - Quantification of pyoverdines concentration
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantification of functional effective interactions
  - Iterative concatenation of FEEs to predict community functions
  - Data selection
  - Data analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.04.016.

ACKNOWLEDGMENTS

We thank S. Kuehn, M. Tikhonov, C. B. Ogbunugafor, M. Rebolleda-Gómez, and all members of the Sanchez lab for helpful discussions. This work was supported by a Packard Foundation Fellowship to A. Sanchez, by the National Institutes of Health through grant 1R35 GM133467-01 to A. Sanchez, and by the Spanish Ministry of Science and Innovation through grant PID2021-125478NA-I00 to A. Sanchez. A. Sanchez and J.D.-C. acknowledge support from grant PID2021-125478NAI00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF: A way of making Europe." A. Sanchez acknowledges funding by the European Union (ERC, ECOPROSPECTOR, 101088469). The views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

AUTHOR CONTRIBUTIONS

D.B. and A. Sanchez conceived the study. J.D.-C. and A. Sanchez designed experiments. J.D.-C. performed experiments. J.D.-C. and A. Skwara processed and analyzed experimental data. J.D.-C. and A. Skwara analyzed data. J.D.-C., A. Skwara, J.C.C.V., D.B., and A. Sanchez discussed and interpreted results. J.D.-C. and A. Sanchez wrote the paper, with input from A. Skwara, J.C.C.V., and D.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 9, 2023
Revised: December 6, 2023
Accepted: April 16, 2024
Published: May 21, 2024

REFERENCES

1. Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science *320*, 1034–1039. https://doi.org/10.1126/science.1153213.
2. Senne de Oliveira Lino, F., Bajic, D., Vila, J.C.C., Sánchez, A., and Sommer, M.O.A. (2021). Complex yeast–bacteria interactions affect the yield of industrial ethanol fermentation. Nat. Commun. *12*, 1498. https://doi.org/10.1038/s41467-021-21844-7.
3. Piccardi, P., Alberti, G., Alexander, J.M., and Mitri, S. (2022). Microbial invasion of a toxic medium is facilitated by a resident community but inhibited as the community co-evolves. ISME J. *16*, 2644–2652. https://doi.org/10.1038/s41396-022-01314-8.
4. Arias-Sanchez, F.I., Vessman, B., Haym, A., Alberti, G., and Mitri, S. (2023). Artificial selection optimizes pollutant-degrading bacterial communities. Preprint at bioRxiv. https://doi.org/10.1101/2023.07.27.550627.
5. Wei, Z., Yang, T., Friman, V.P., Xu, Y., Shen, Q., and Jousset, A. (2015). Trophic network architecture of root-associated bacterial communities determines pathogen invasion and plant health. Nat. Commun. *6*, 8413. https://doi.org/10.1038/ncomms9413.
6. Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., and Cooper, T.F. (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. Science *332*, 1193–1196. https://doi.org/10.1126/science.1203801.
7. Sanchez, A., Bajic, D., Diaz-Colunga, J., Skwara, A., Vila, J.C.C., and Kuehn, S. (2023). The community-function landscape of microbial consortia. Cell Syst. *14*, 122–134. https://doi.org/10.1016/j.cels.2022.12.011.
8. Chen, Y., Lin, C.J., Jones, G., Fu, S., and Zhan, H. (2009). Enhancing biodegradation of wastewater by microbial consortia with fractional factorial design. J. Hazard. Mater. *171*, 948–953. https://doi.org/10.1016/j.jhazmat.2009.06.100.
9. Eng, A., and Borenstein, E. (2019). Microbial community design: methods, applications, and opportunities. Curr. Opin. Biotechnol. *58*, 117–128. https://doi.org/10.1016/j.copbio.2019.03.002.
10. Gowda, K., Ping, D., Mani, M., and Kuehn, S. (2022). Genomic structure predicts metabolite dynamics in microbial communities. Cell *185*, 530–546.e25. https://doi.org/10.1016/j.cell.2021.12.036.
11. Thompson, J., Johansen, R., Dunbar, J., and Munsky, B. (2019). Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS One *14*, e0215502. https://doi.org/10.1371/journal.pone.0215502.
12. Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of machine learning in microbiology. Front. Microbiol. *10*, 827. https://doi.org/10.3389/fmicb.2019.00827.
13. Clark, R.L., Connors, B.M., Stevenson, D.M., Hromada, S.E., Hamilton, J.J., Amador-Noguez, D., and Venturelli, O.S. (2021). Design of synthetic human gut microbiome assembly and butyrate production. Nat. Commun. *12*, 3254. https://doi.org/10.1038/s41467-021-22938-y.
14. Skwara, A., Gowda, K., Yousef, M., Diaz-Colunga, J., Raman, A.S., Sanchez, A., Tikhonov, M., and Kuehn, S. (2023). Statistically learning the functional landscape of microbial communities. Nat. Ecol. Evol. *7*, 1823–1833. https://doi.org/10.1038/s41559-023-02197-4.
15. Cardinale, B.J., Palmer, M.A., and Collins, S.L. (2002). Species diversity enhances ecosystem functioning through interspecific facilitation. Nature *415*, 426–429. https://doi.org/10.1038/415426a.
16. Bell, T., Newman, J.A., Silverman, B.W., Turner, S.L., and Lilley, A.K. (2005). The contribution of species richness and composition to bacterial services. Nature *436*, 1157–1160. https://doi.org/10.1038/nature03891.
17. Loreau, M., and Hector, A. (2001). Partitioning selection and complementarity in biodiversity experiments. Nature *412*, 72–76. https://doi.org/10.1038/35083573.
18. Kirwan, L., Connolly, J., Finn, J.A., Brophy, C., Lüscher, A., Nyfeler, D., and Sebastià, M.T. (2009). Diversity–interaction modeling: estimating contributions of species identities and interactions to ecosystem function. Ecology *90*, 2032–2038. https://doi.org/10.1890/08-1684.1.
19. Connolly, J., Bell, T., Bolger, T., Brophy, C., Carnus, T., Finn, J.A., Kirwan, L., Isbell, F., Levine, J., Lüscher, A., et al. (2013). An improved model to predict the effects of changing biodiversity levels on ecosystem function. J. Ecol. *101*, 344–355. https://doi.org/10.1111/1365-2745.12052.
20. Wagg, C., Bender, S.F., Widmer, F., and Van Der Heijden, M.G.A. (2014). Soil biodiversity and soil community composition determine ecosystem multifunctionality. Proc. Natl. Acad. Sci. USA *111*, 5266–5270. https://doi.org/10.1073/pnas.1320054111.
21. Midgley, G.F. (2012). Ecology. Biodiversity and ecosystem function. Science *335*, 174–175. https://doi.org/10.1126/science.1217245.
22. Bank, C. (2022). Epistasis and adaptation on fitness landscapes. Annu. Rev. Ecol. Evol. Syst. *53*, 457–479. https://doi.org/10.1146/annurev-ecolsys-102320-112153.
23. Taylor, M.B., and Ehrenreich, I.M. (2015). Higher-order genetic interactions and their contribution to complex traits. Trends Genet. *31*, 34–40. https://doi.org/10.1016/j.tig.2014.09.001.
24. Sailer, Z.R., and Harms, M.J. (2017). Detecting high-order epistasis in nonlinear genotype-phenotype maps. Genetics *205*, 1079–1088. https://doi.org/10.1534/genetics.116.195214.
25. Sailer, Z.R., and Harms, M.J. (2017). High-order epistasis shapes evolutionary trajectories. PLoS Comput. Biol. *13*, e1005541. https://doi.org/10.1371/journal.pcbi.1005541.
26. Weinreich, D.M., Lan, Y., Jaffe, J., and Heckendorn, R.B. (2018). The influence of higher-order epistasis on biological fitness landscape topography. J. Stat. Phys. *172*, 208–225. https://doi.org/10.1007/s10955-018-1975-3.
27. Bank, C., Matuszewski, S., Hietpas, R.T., and Jensen, J.D. (2016). On the (un)predictability of a large intragenic fitness landscape. Proc. Natl. Acad. Sci. USA *113*, 14085–14090. https://doi.org/10.1073/pnas.1612676113.
28. Yang, G., Anderson, D.W., Baier, F., Dohmen, E., Hong, N., Carr, P.D., Kamerlin, S.C.L., Jackson, C.J., Bornberg-Bauer, E., and Tokuriki, N. (2019). Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. Nat. Chem. Biol. *15*, 1120–1128. https://doi.org/10.1038/s41589-019-0386-3.
29. MacLean, R.C., Perron, G.G., and Gardner, A. (2010). Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in Pseudomonas aeruginosa. Genetics *186*, 1345–1354. https://doi.org/10.1534/genetics.110.123083.
30. Chou, H.H., Chiu, H.C., Delaney, N.F., Segrè, D., and Marx, C.J. (2011). Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science *332*, 1190–1192. https://doi.org/10.1126/science.1203799.
31. Perfeito, L., Sousa, A., Bataillon, T., and Gordo, I. (2014). Rates of fitness decline and rebound suggest pervasive epistasis. Evolution *68*, 150–162. https://doi.org/10.1111/evo.12234.

32. Kryazhimskiy, S., Rice, D.P., Jerison, E.R., and Desai, M.M. (2014). Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science *344*, 1519–1522. https://doi.org/10.1126/science.1250939.

33. Schoustra, S., Hwang, S., Krug, J., and de Visser, J.A.G. (2016). Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. Proc. Biol. Sci. *283*, 20161376. https://doi.org/10.1098/rspb.2016.1376.

34. Otwinowski, J., McCandlish, D.M., and Plotkin, J.B. (2018). Inferring the shape of global epistasis. Proc. Natl. Acad. Sci. USA *115*, E7550–E7558. https://doi.org/10.1073/pnas.1804015115.

35. Johnson, M.S., Martsul, A., Kryazhimskiy, S., and Desai, M.M. (2019). Higher-fitness yeast genotypes are less robust to deleterious mutations. Science *366*, 490–493. https://doi.org/10.1126/science.aay4199.

36. Wei, X., and Zhang, J. (2019). Patterns and mechanisms of diminishing returns from beneficial mutations. Mol. Biol. Evol. *36*, 1008–1021. https://doi.org/10.1093/molbev/msz035.

37. Bakerlee, C.W., Nguyen Ba, A.N., Shulgina, Y., Rojas Echenique, J.I., and Desai, M.M. (2022). Idiosyncratic epistasis leads to global fitness–correlated trends. Science *376*, 630–635. https://doi.org/10.1126/science.abm4774.

38. Reddy, G., and Desai, M.M. (2021). Global epistasis emerges from a generic model of a complex trait. eLife *10*, e64740. https://doi.org/10.7554/eLife.64740.

39. Diaz-Colunga, J., Skwara, A., Gowda, K., Diaz-Uriarte, R., Tikhonov, M., Bajic, D., and Sanchez, A. (2023). Global epistasis on fitness landscapes. Philos. Trans. R. Soc. Lond. B Biol. Sci. *378*, 20220053. https://doi.org/10.1098/rstb.2022.0053.

40. Johnson, M.S., Reddy, G., and Desai, M.M. (2023). Epistasis and evolution: recent advances and an outlook for prediction. BMC Biol. *21*, 120. https://doi.org/10.1186/s12915-023-01585-3.

41. Romero, P.A., Krause, A., and Arnold, F.H. (2013). Navigating the protein fitness landscape with Gaussian processes. Proc. Natl. Acad. Sci. USA *110*, E193–E201. https://doi.org/10.1073/pnas.1215251110.

42. Tareen, A., Kooshkbaghi, M., Posfai, A., Ireland, W.T., McCandlish, D.M., and Kinney, J.B. (2022). MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. Genome Biol. *23*, 98. https://doi.org/10.1186/s13059-022-02661-7.

43. Otwinowski, J. (2018). Biophysical inference of epistasis and the effects of mutations on protein stability and function. Mol. Biol. Evol. *35*, 2345–2354. https://doi.org/10.1093/molbev/msy141.

44. Sailer, Z.R., Shafik, S.H., Summers, R.L., Joule, A., Patterson-Robert, A., Martin, R.E., and Harms, M.J. (2020). Inferring a complete genotype-phenotype map from a small number of measured phenotypes. PLoS Comput. Biol. *16*, e1008243. https://doi.org/10.1371/journal.pcbi.1008243.

45. Tonner, P.D., Pressman, A., and Ross, D. (2022). Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. Proc. Natl. Acad. Sci. USA *119*, e2114021119. https://doi.org/10.1073/pnas.2114021119.

46. Gould, A.L., Zhang, V., Lamberti, L., Jones, E.W., Obadia, B., Korasidis, N., Gavryushkin, A., Carlson, J.M., Beerenwinkel, N., and Ludington, W.B. (2018). Microbiome interactions shape host fitness. Proc. Natl. Acad. Sci. USA *115*, E11951–E11960. https://doi.org/10.1073/pnas.1809349115.

47. Sanchez-Gorostiaga, A., Bajić, D., Osborne, M.L., Poyatos, J.F., and Sanchez, A. (2019). High-order interactions distort the functional landscape of microbial consortia. PLoS Biol. *17*, e3000550. https://doi.org/10.1371/journal.pbio.3000550.

48. Morris, A., Meyer, K., and Bohannan, B. (2020). Linking microbial communities to ecosystem functions: what we can learn from genotype–phenotype mapping in organisms. Philos. Trans. R. Soc. Lond. B Biol. Sci. *375*, 20190244. https://doi.org/10.1098/rstb.2019.0244.

49. Eble, H., Joswig, M., Lamberti, L., and Ludington, W.B. (2021). High dimensional geometry of fitness landscapes identifies master regulators of evolution and the microbiome. Preprint at bioRxiv. https://doi.org/10.1073/pnas.2300634120.

50. Mould, D.L., Botelho, N.J., and Hogan, D.A. (2020). Intraspecies signaling between common variants of Pseudomonas aeruginosa increases production of quorum-sensing-controlled virulence factors. mBio *11*, e01865–20. https://doi.org/10.1128/mBio.01865-20.

51. Stintzi, A., Evans, K., Meyer, J.M., and Poole, K. (1998). Quorum-sensing and siderophore biosynthesis in Pseudomonas aeruginosa: lasRllasI mutants exhibit reduced pyoverdine biosynthesis. FEMS Microbiol. Lett. *166*, 341–345. https://doi.org/10.1111/j.1574-6968.1998.tb13910.x.

52. Langenheder, S., Bulling, M.T., Solan, M., and Prosser, J.I. (2010). Bacterial biodiversity-ecosystem functioning relations are modified by environmental complexity. PLoS One *5*, e10834. https://doi.org/10.1371/journal.pone.0010834.

53. Kuebbing, S.E., Classen, A.T., Sanders, N.J., and Simberloff, D. (2015). Above-and below-ground effects of plant diversity depend on species origin: an experimental test with multiple invaders. New Phytol. *208*, 727–735. https://doi.org/10.1111/nph.13488.

54. Ghedini, G., Marshall, D.J., and Loreau, M. (2022). Phytoplankton diversity affects biomass and energy production differently during community development. Funct. Ecol. *36*, 446–457. https://doi.org/10.1111/1365-2435.13955.

55. Berger, D., and Postma, E. (2014). Biased estimates of diminishing-returns epistasis? Empirical evidence revisited. Genetics *198*, 1417–1420. https://doi.org/10.1534/genetics.114.169870.

56. Ruiz, J., de Celis, M., Diaz-Colunga, J., Vila, J.C., Benitez-Dominguez, B., Vicente, J., Santos, A., Sanchez, A., and Belda, I. (2023). Predictability of the community-function landscape in wine yeast ecosystems. Mol. Syst. Biol. *19*, e11613. https://doi.org/10.15252/msb.202311613.

57. Shade, A. (2017). Diversity is the question, not the answer. ISME J. *11*, 1–6. https://doi.org/10.1038/ismej.2016.118.

58. Billick, I., and Case, T.J. (1994). Higher order interactions in ecological communities: what are they and how can they be detected? Ecology *75*, 1529–1543. https://doi.org/10.2307/1939614.

59. Guo, X., and Boedicker, J. (2016). High-order interactions between species strongly influence the activity of microbial communities. Biophys. J. *110*, 143a. https://doi.org/10.1016/j.bpj.2015.11.811.

60. Letten, A.D., and Stouffer, D.B. (2019). The mechanistic basis for higher-order interactions and non-additivity in competitive communities. Ecol. Lett. *22*, 423–436. https://doi.org/10.1111/ele.13211.

61. Mickalide, H., and Kuehn, S. (2019). Higher-order interaction between species inhibits bacterial invasion of a phototroph-predator microbial community. Cell Syst. *9*, 521–533.e10. https://doi.org/10.1016/j.cels.2019.11.004.

62. Korkmazhan, E., and Dunn, A.R. (2022). High-order correlations in species interactions lead to complex diversity-stability relationships for ecosystems. Phys. Rev. E *105*, 014406. https://doi.org/10.1103/PhysRevE.105.014406.

63. Ghoul, M., and Mitri, S. (2016). The ecology and evolution of microbial competition. Trends Microbiol. *24*, 833–845. https://doi.org/10.1016/j.tim.2016.06.011.

64. Amor, D.R., Ratzke, C., and Gore, J. (2020). Transient invaders can induce shifts between alternative stable states of microbial communities. Sci. Adv. *6*, eaay8676. https://doi.org/10.1126/sciadv.aay8676.

65. Daniels, M., van Vliet, S., and Ackermann, M. (2023). Changes in interactions over ecological time scales influence single-cell growth dynamics in a metabolically coupled marine microbial community. ISME J. *17*, 406–416. https://doi.org/10.1038/s41396-022-01312-w.

66. Arya, S., George, A.B., and O'Dwyer, J.P. (2023). Sparsity of higher-order landscape interactions enables learning and prediction for microbiomes.

Proc. Natl. Acad. Sci. USA *120*, e2307313120. https://doi.org/10.1073/pnas.2307313120.

67. Sun, X., Folmar, J., Favier, A., Pyenson, N., Sanchez, A., and Rebolleda-Gomez, M. (2023). Predictive microbial community changes across a temperature gradient. Preprint at bioRxiv. https://doi.org/10.1101/2023.07.28.550899.
68. Ratzke, C., and Gore, J. (2018). Modifying and reacting to the environmental pH can drive bacterial interactions. PLOS Biol. *16*, e2004248. https://doi.org/10.1371/journal.pbio.2004248.
69. Hu, J., Amor, D.R., Barbier, M., Bunin, G., and Gore, J. (2022). Emergent phases of ecological diversity and dynamics mapped in microcosms. Science *378*, 85–89. https://doi.org/10.1126/science.abm7841.
70. Diaz-Colunga, J., Sanchez, A., and Ogbunugafor, C.B. (2023). Environmental modulation of global epistasis in a drug resistance fitness landscape. Nat. Commun. *14*, 8055. https://doi.org/10.1038/s41467-023-43806-x.
71. Ardell, S., Martsul, A., Johnson, M.S., and Kryazhimskiy, S. (2023). Environment-independent distribution of mutational effects emerges from microscopic epistasis. Preprint at bioRxiv. https://doi.org/10.1101/2023.11.18.567655.
72. Diaz-Colunga, J., Lu, N., Sanchez-Gorostiaga, A., Chang, C.Y., Cai, H.S., Goldford, J.E., Tikhonov, M., and Sánchez, Á. (2022). Top-down and bottom-up cohesiveness in microbial community coalescence. Proc. Natl. Acad. Sci. USA *119*, e2111261119. https://doi.org/10.1073/pnas.2111261119.
73. Drake, E.J., Cao, J., Qu, J., Shah, M.B., Straubinger, R.M., and Gulick, A.M. (2007). The 1.8 Å crystal structure of PA2412, an MbtH-like protein from the pyoverdine cluster of Pseudomonas aeruginosa. J. Biol. Chem. *282*, 20425–20434. https://doi.org/10.1074/jbc.M611833200.
74. George, A.B., and Korolev, K.S. (2023). Ecological landscapes guide the assembly of optimal microbial communities. PLoS Comput. Biol. *19*, e1010570. https://doi.org/10.1371/journal.pcbi.1010570.
75. R Core Team (2022). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Bacterial strains used for assembling pyoverdine-producing consortia (*Pseudomonas* sp. 01 to 05, *Enterobacter* sp., *Klebsiella* sp., *Raoultella* sp.)	Diaz-Colunga et al.[72]	N/A
Chemicals, peptides, and recombinant proteins		
Sodium phosphate dihydrate ($Na_2HPO_4 \times 2H_2O$)	Sigma-Aldrich	Cat# 04269
Potassium phosphate (KH_2PO_4)	Fisher Scientific	Cat# P285-500
Ammonium chloride (NH_4Cl)	Fisher Scientific	Cat# A661-500
Sodium chloride (NaCl)	Fisher Scientific	Cat# S271-500
Sodium citrate	Sigma-Aldrich	Cat# S4641
Calcium chloride ($CaCl_2$)	Sigma-Aldrich	Cat# 102392
Magnesium sulfate ($MgSO_4$)	Fisher Scientific	Cat# M65-500
Trace mineral supplement	ATCC	Cat# MD-TMS
Deposited data		
Data generated for this study (composition and function of pyoverdine-producing consortia)	GitHub	https://github.com/jdiazc9/eco_global_epist
Plant biomass dataset	Kuebbing et al.[53]	N/A
Phytoplankton biomass dataset	Ghedini et al.[54]	N/A
Xylose oxidation dataset	Langenheder et al.[52]	N/A
Starch hydrolysis dataset	Sanchez-Gorostiaga et al.[47]	N/A
Butyrate secretion dataset	Clark et al.[13]	N/A
Software and algorithms		
Code used for the analysis and figures	GitHub	https://github.com/jdiazc9/eco_global_epist

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Alvaro Sanchez (alvaro.sanchez@usal.es).

Materials availability

This study did not generate new unique materials.

Data and code availability

All original data and code is publicly available via the GitHub repository: https://github.com/jdiazc9/eco_global_epist. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Bacterial species isolation from environmental samples

Bacterial isolates were obtained from the communities described in Diaz-Colunga et al.[72] In short, environmental samples (soil, leaves...) were collected from different geographical locations and used to inoculate eight identical synthetic habitats containing 500 μL of minimal medium with citrate as the only supplied carbon source. Communities were stabilized by serial passaging for 7 cycles of 48 h growth and 1:125 dilution such that species unable to grow in these conditions were filtered out. Communities

were then diluted to 10^{-6} and streaked out on chromogenic agar plates (CHROMagar Mastitis GN). Eight individual colonies were selected such that they were distinguishable from one another in their color and morphology. Isolates were streaked out two more times to filter out potential contamination from different species. Isolates were identified at the genus level via Sanger sequencing of the 16S rRNA gene (Table S2) — all 16S sequences were less than 97% similar.

METHOD DETAILS

Combinatorial community assembly

Experiments were done using M9 minimal medium containing 4.65 g/L $Na_2HPO_4 \times 2H_2O$ (Sigma-Aldrich), 3 g/L KH_2PO_4 (Fisher Scientific), 1 g/L NH_4Cl (Fisher Scientific), 0.5 g/L NaCl (Fisher Scientific) supplemented with 2.5 g/L sodium citrate (Sigma-Aldrich), 0.1 mM $CaCl_2$ (Sigma-Aldrich), 2 mM $MgSO_4$ (Fisher Scientific) and 1% trace mineral supplement (v/v; ATCC; stock contains 0.5 g/L Ethylenediaminetetraacetic acid [EDTA], 3 g/L $MgSO_4 \times 7H_2O$, 0.5 g/L $MnSO_4 \times H2O$, 1 g/L NaCl, 0.1 g/L $FeSO_4 \times 7H2O$, 0.1 g/L $Co[NO_3]_2 \times 6H_2O$, 0.1 g/L $CaCl_2$ [anhydrous], 0.1 g/L $ZnSO_4 \times 7H_2O$, 0.01 g/L $CuSO_4 \times 5H_2O$, 0.01 g/L $AlK[SO_4]_2$ [anhydrous], 0.01 g/L H_3BO_3, 0.01 g/L $Na_2MoO_4 \times 2H_2O$, 0.001 g/L Na_2SeO_3 [anhydrous], 0.01 g/L $Na_2WO_4 \times 2H_2O$ and 0.02 g/L $NiCl_2 \times 6H_2O$). Starter cultures were prepared by resuspending a single colony of each isolate into individual 50 mL conical tubes (Falcon) containing 20 mL of medium, and allowing them to grow for 48 h at 30°C. Cultures were then fully homogenized, and communities were assembled in 96-deep well U-bottom plates (Greiner Bio-One) filled with 500 μL of medium per well by inoculating 1 μL of each starter monoculture into the corresponding wells — further details can be found in the Methods S1. Communities were then incubated still at 30°C for 48 h. The experiment was independently replicated three times, by the same researcher on different days.

Quantification of pyoverdines concentration

After incubation, cultures were fully homogenized and growth was tracked by measuring the optical density (OD) at 620 nm of 100 μL of culture in an AccuSkan FC plate reader (Fisher Scientific). Cells were then pelleted by centrifuging the plates at 3000 rpm for 25 min. Supernatants were collected and filtered through multi-well 0.2 μm filters (Pall Corporation) by centrifuging at 3500 rpm for 5 min. The OD at 405 nm of 100 μL of the supernatants was quantified in the same plate reader. OD values were converted to units of concentration using a value of 1.9×10^4 $M^{-1}cm^{-1}$ for the extinction coefficient of pyoverdines at 405 nm.[73]

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of functional effective interactions

For each dataset we analyzed, the effective interaction between species j and j ($\tilde{\epsilon}_{ij}$) was quantified as explained in Figures 3A–3C and in the Methods S1, that is:

$$\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \frac{\langle \Delta F_j \rangle_{B(i)}}{\sum_{k \neq i} \langle \Delta F_k \rangle^2_{B(i)}} \quad \text{(Equation 3)}$$

The term ϵ_{ij} represents the deviation between the function of a community that contains both species i and j with respect to the additive expectation that they do not interact (Figure 3B), that is, that the difference in function between communities $\mathbf{s}$ and $\mathbf{s}+\mathbf{i}+\mathbf{j}$ is the sum of the separate contributions of species i and j. Mathematically, this means:

$$\epsilon_{ij} = \underbrace{F(\mathbf{s}+\mathbf{i}+\mathbf{j})}_{\substack{\text{function of community}\\ \text{containing both}\\ \text{species } i \text{ and } j}} - \underbrace{[F(\mathbf{s})+\Delta F_i(\mathbf{s})+\Delta F_j(\mathbf{s})]}_{\substack{\text{additive}\\ \text{expectation}}} \quad \text{(Equation 4)}$$

$$\epsilon_{ij} = F(\mathbf{s}+\mathbf{i}+\mathbf{j}) - \left[F(\mathbf{s}) + \underbrace{[F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s})]}_{\Delta F_i(\mathbf{s})} + \underbrace{[F(\mathbf{s}+\mathbf{j}) - F(\mathbf{s})]}_{\Delta F_j(\mathbf{s})}\right] \quad \text{(Equation 5)}$$

$$\epsilon_{ij} = F(\mathbf{s}+\mathbf{i}+\mathbf{j}) - F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s}+\mathbf{j}) + F(\mathbf{s}) \quad \text{(Equation 6)}$$

To quantify the average $\langle \epsilon_{ij} \rangle$ as it appears in Equation 3, we applied Equation 6 to every possible background community $\mathbf{s}$ not containing species i nor j, and took the average over all ϵ_{ij}. Note that not every dataset we analyzed is combinatorially complete, so in many instances some of the terms in Equation 6 might be unknown. In those cases, we computed the average over only the known values of ϵ_{ij}.

The terms ΔF_j in Equation 3 represent the functional effect of species j as defined in the main text, calculated simply as $F(\mathbf{s}+\mathbf{j}) - F(\mathbf{s})$. In Equation 3, these values appear averaged across those background communities that do not contain species i (nor,

naturally, species j) — this set of backgrounds is denoted as $B(i)$. Like before, whenever there were unknown terms due to incomplete data we averaged across the known values only.

Iterative concatenation of FEEs to predict community functions

We call $\mathbf{s}_0$ one of the empirically tested consortia (henceforth an in-sample community). One may then want to predict the function of an out-of-sample consortium (which we denote as $\mathbf{s}_1$). In the example shown in Figure 4A, the out-of-sample consortium contains three more species (i, j, and k) than the in-sample one. Including the first species i in the starting consortium $\mathbf{s}_0$ will have an effect in function that we can estimate from the linear FEE for species i: $\Delta F_i(\mathbf{s}_0) = a_i + b_i F(\mathbf{s}_0)$. The function of the consortium resulting from the addition of species i to $\mathbf{s}_0$ would then simply be $F(\mathbf{s}_0 + \mathbf{i}) = F(\mathbf{s}_0) + \Delta F_i(\mathbf{s}_0) = a_i + (1 + b_i)F(\mathbf{s}_0)$. We can next estimate the functional effect of including species j on the "updated" background consortium $\mathbf{s}_0 + \mathbf{i}$, and finally the functional effect of species k on $\mathbf{s}_0 + \mathbf{i} + \mathbf{j}$ (Figure 4A):

$$F(\mathbf{s}_1) = \underbrace{F(\mathbf{s}_0)}_{\substack{\text{starting in-sample}\\ \text{community function}}} + \underbrace{\Delta F_i(\mathbf{s}_0)}_{\substack{\text{functional effect of}\\ \text{species } i\\ \text{on community } \mathbf{s}_0}} + \underbrace{\Delta F_j(\mathbf{s}_0+\mathbf{i})}_{\substack{\text{functional effect of}\\ \text{species } j\\ \text{on community } \mathbf{s}_0+\mathbf{i}}} + \underbrace{\Delta F_k(\mathbf{s}_0+\mathbf{i}+\mathbf{j})}_{\substack{\text{functional effect of}\\ \text{species } k\\ \text{on community } \mathbf{s}_0+\mathbf{i}+\mathbf{j}}} \quad \text{(Equation 7)}$$

This iterative procedure ultimately gives a prediction for the out-of-sample community function $F(\mathbf{s}_1)$. By estimating the residuals of the FEEs, we can refine predictions and ensure that the order of species addition (e.g., $i \rightarrow j \rightarrow k$, or $k \rightarrow i \rightarrow j$, or $j \rightarrow i \rightarrow k$, etc.) does not affect the predicted value — this is further discussed in the Methods S1.

Data selection

The datasets re-analyzed in this work have been previously used in studies aiming to connect the species-level composition with the function of ecological communities.[14,74] We re-analyzed all datasets we could find which were peer-reviewed and publicly available, where a quantitative function had been mapped to the presence or absence of each species, where four or more species were considered, and where data were sufficiently complete so that each FEE could be fit to at least five data points.

Data analysis

All analyses were performed using R version 4.3.1.[75]

Cell
Theory

CellPress
OPEN ACCESS

Supplemental figures

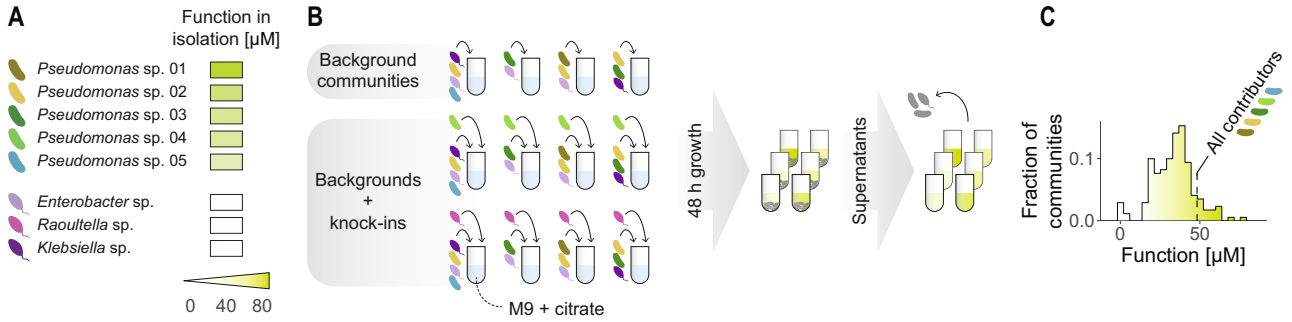


Figure S1. Assembly of microbial consortia in synthetic laboratory conditions, related to Figure 1

(A) We isolated eight bacterial species from environmental samples and identified them at the genus level (STAR Methods). Five of them exhibited secretion of pyoverdines when grown in monoculture in minimal M9 citrate medium (STAR Methods).

(B) We assembled 164 consortia by inoculating combinations of these eight species into minimal M9 citrate medium and incubated them for 48 h. We then collected the spent media and quantified the concentration of pyoverdines in them (STAR Methods).

(C) We found variable levels of pyoverdines secretion, with the concentrations in the supernatants ranging from 0 to roughly 70 μM. About 20% of the assemblages exhibited higher function than the consortium formed by all five pyoverdines secretors.

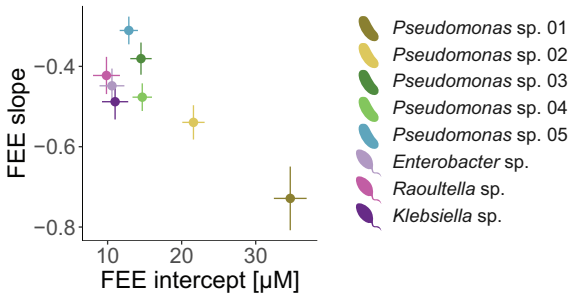


Figure S2. FEE slopes and intercepts vary across species, related to Figure 1

Each species in our pyoverdine experiment exhibits a different FEE, characterized by its slope and intercept. Error bars represent standard deviations of the linear fit coefficients in Figure 1F of the main text.

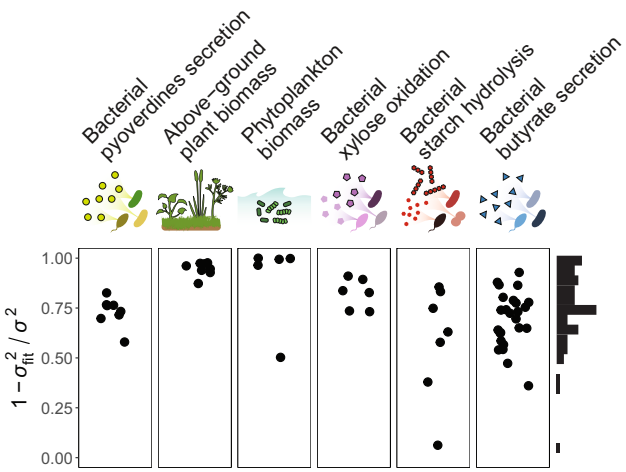


Figure S3. FEEs explain a high fraction of the variation in the functional effect of a species, related to Figures 1 and 2

We fit a linear functional effect equation (FEE) to every species in each of the datasets described in the main text. The quality of the fits is here quantified as $1 - \sigma^2_{fit}/\sigma^2$, where σ^2 is the variance of the distribution of functional effects across all species in the dataset, and σ^2_{fit} is the variance of the residuals of the fit for a given species. This is a metric of the degree to which the functional effect of a species is predictable from its FEE.

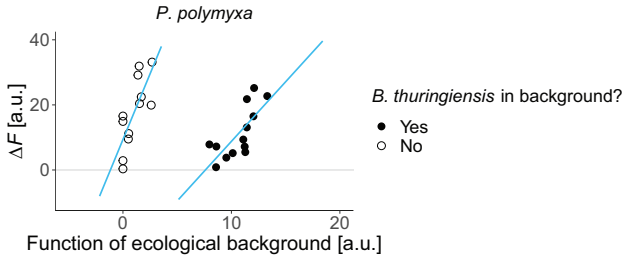


Figure S4. Branching in the functional effect equation of a species, related to Figure 2

The effect on the amylolytic activity[47] of a community induced by *P. polymyxa* follows a different scaling with the function of the ecological background depending on whether *B. thuringiensis* is present (filled dots) or absent (hollow dots) in the background.

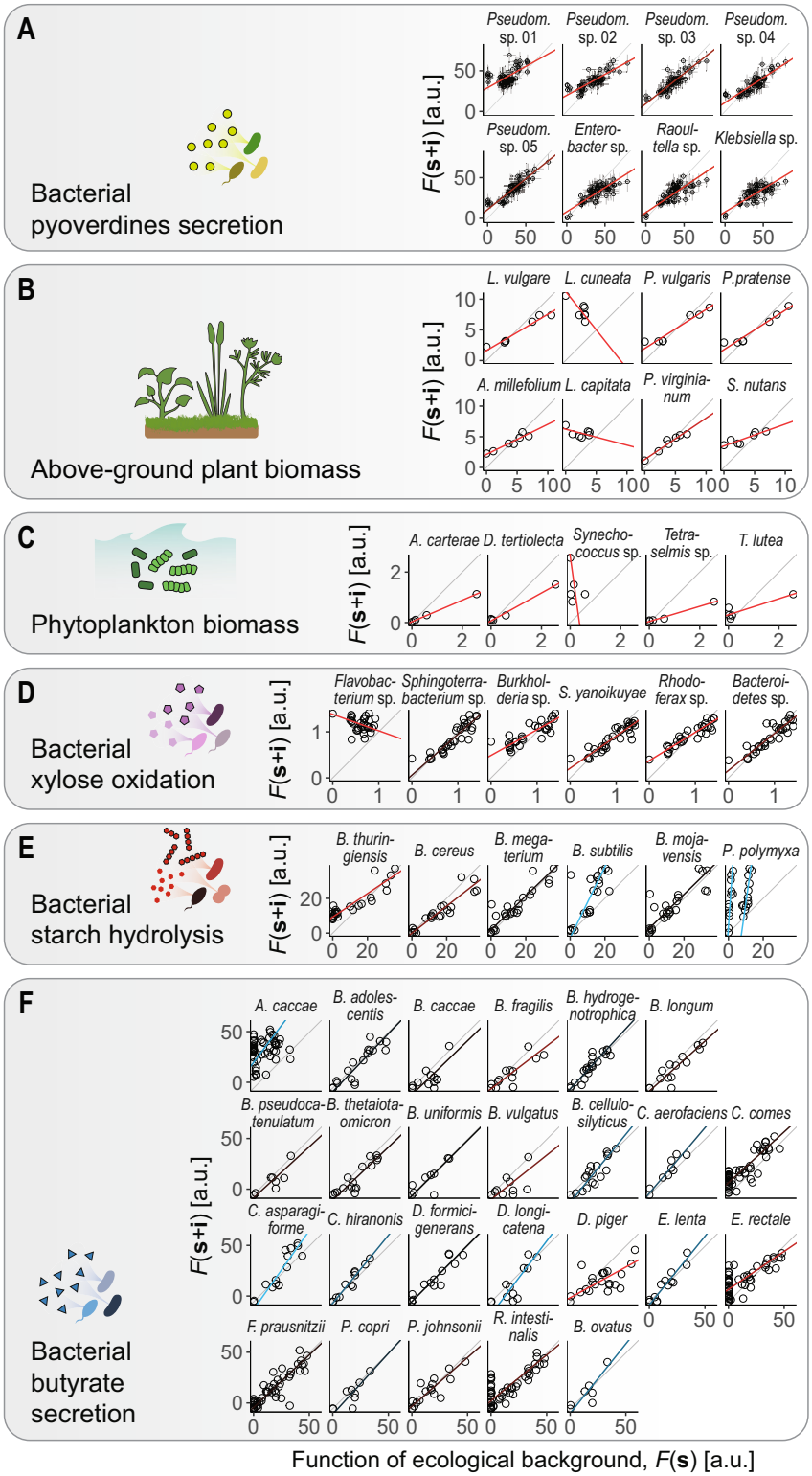


Figure S5. Ecological function with and without a focal species across datasets, related to Figures 1 and 2

Representing the functional effect ΔF versus the background function $F(\mathbf{s})$ (as we do in Figures 1F and 2 in the main text) may lead to biased estimates of the degree of global epistasis since ΔF explicitly depends on $F(\mathbf{s})$: $\Delta F = F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s})$.[55] Here we represent the function of the background consortium, $F(\mathbf{s})$,

(legend continued on next page)

directly against the function of the consortium where the focal species i has been added, $F(\mathbf{s}+\mathbf{i})$, for every focal species across datasets. Colored lines correspond to total least squares regressions to the data (blue: slope greater than 1; red: slope lower than 1). Gray lines indicate $y = x$. Data sources:
(A) This study. Dots and error bars represent means and standard deviations across three biological replicates.
(B) Kuebbing et al.[53]
(C) Ghedini et al.[54]
(D) Langenheder et al.[52]
(E) Sanchez-Gorostiaga et al.[47]
(F) Clark et al.[13]

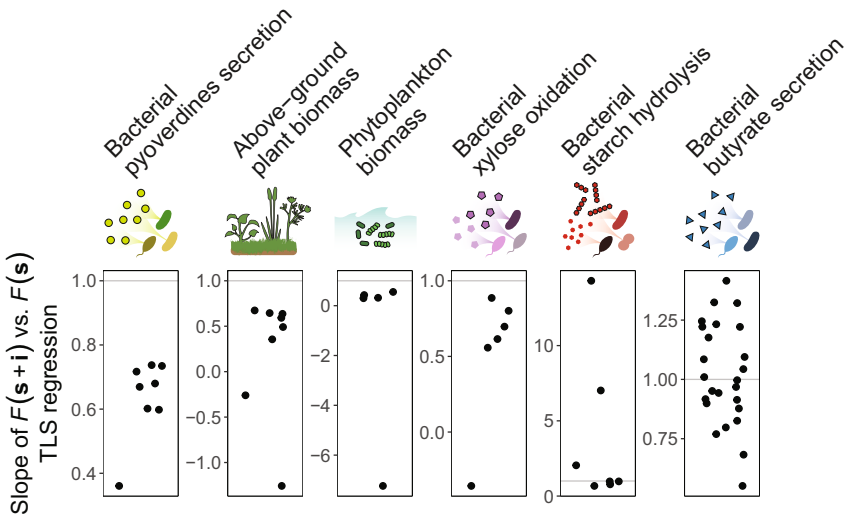


Figure S6. Slopes of $F(s+i)$ versus $F(s)$ regressions across datasets, related to Figures 1 and 2

We represent the slope of the total least squares (TLS) regressions to $F(\mathbf{s}+\mathbf{i})$ versus $F(\mathbf{s})$ (shown in Figure S5) for all species across datasets. Gray horizontal lines correspond to a slope of 1. Recent work in genetics[37] has quantified the degree of global epistasis by studying whether the slopes of such TLS regressions differ from 1.

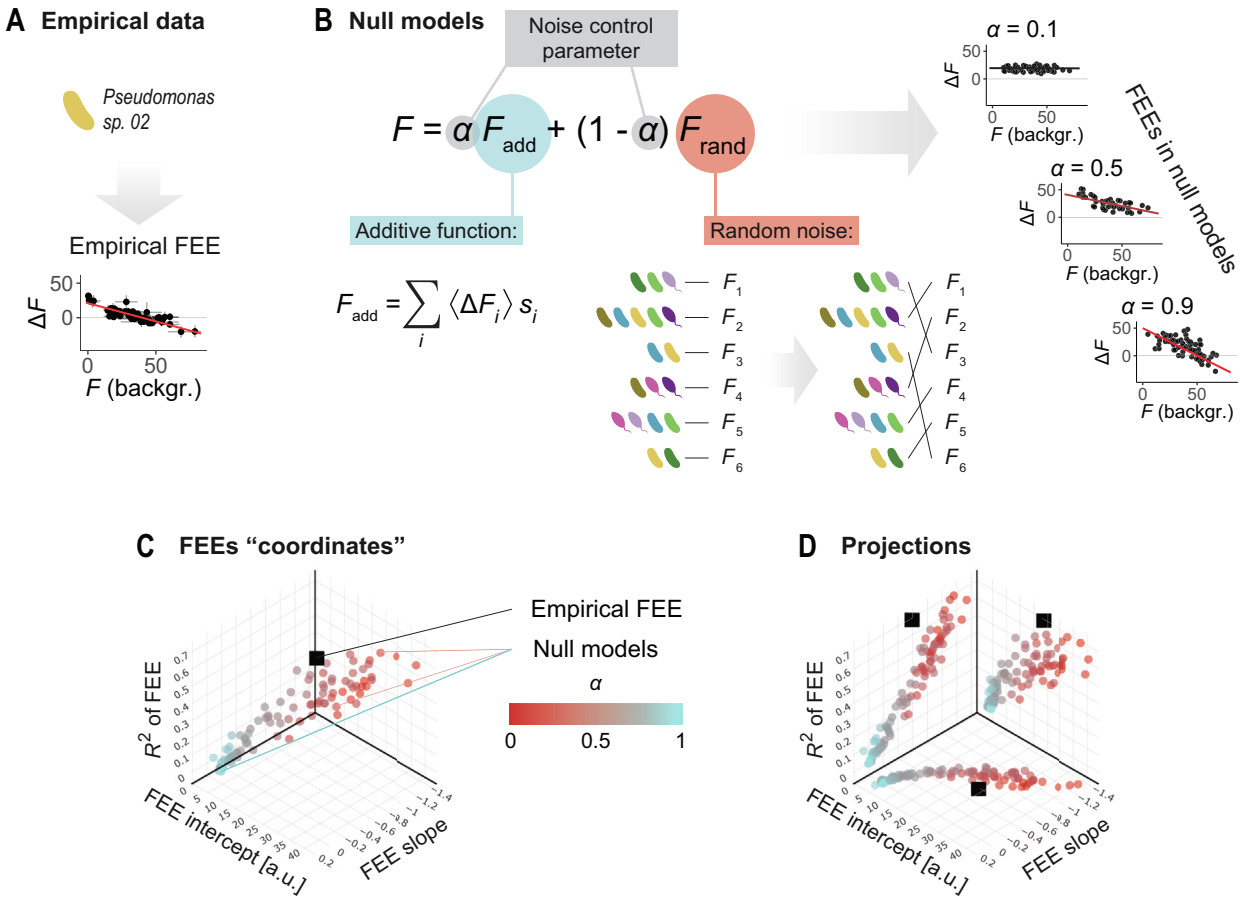


Figure S7. Comparison between null models and empirical FEE for *Pseudomonas* sp. 02, related to Figures 1 and 2

The empirical FEE observed for *Pseudomonas* sp. 02 is not compatible with a null model.

(A) Empirical FEE for *Pseudomonas* sp. 02 (see Figure 1F in main text).

(B) We consider a family of null models where the function F of a consortium is given by a composition of an additive contribution plus random noise (see Methods S1 for details). The parameter α controls the amount of noise ($\alpha = 1$ corresponds to the strictly additive case, $\alpha = 0$ corresponds to the case where F is dominated by noise and the mapping between community composition and function becomes random). Different FEEs emerge within this family of null models.

(C) We quantified the slope, intercept, and R^2 of the FEEs that emerged in 1,000 null models (only 100 are represented here for visual clarity) with different values for α (colored dots). We compare these quantities with those corresponding to the empirical FEE for *Pseudomonas* sp. 02 (black square).

(D) Same as previous panel, but here we represent the projections on the x-y, x-z, and y-z planes.

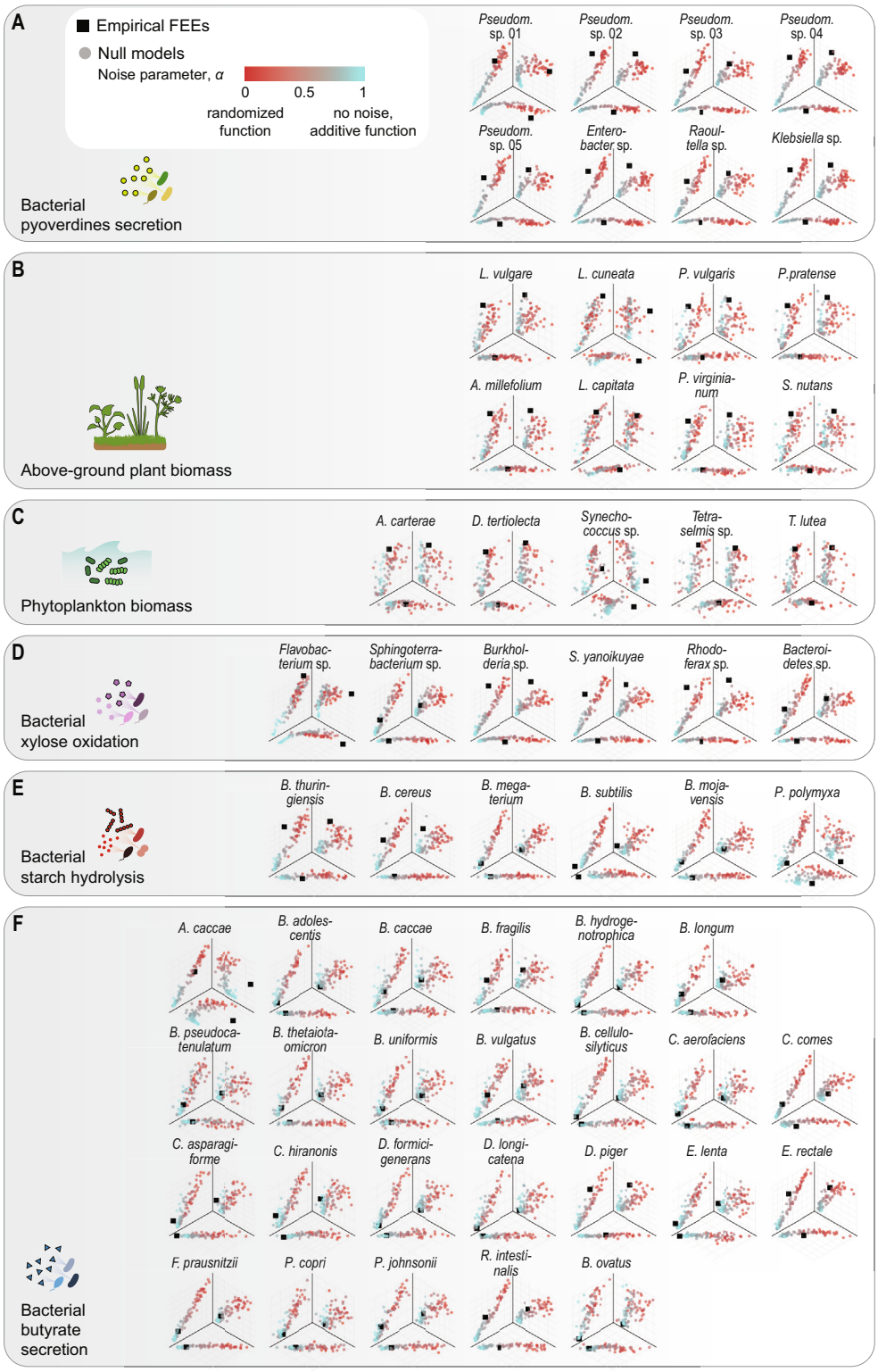


Figure S8. Empirically observed FEEs are often not compatible with a null model, related to Figures 1 and 2

We repeated the analysis shown in Figure S7D for every species and dataset. Here we plot the projections in the 3-dimensional space defined by the FEE slope, intercept, and R^2, of the FEEs corresponding to the null models (colored dots) and empirical FEEs (black squares). Data sources:
(A) This study.

(legend continued on next page)

(B) Kuebbing et al.[53]
(C) Ghedini et al.[54]
(D) Langenheder et al.[52]
(E) Sanchez-Gorostiaga et al.[47]
(F) Clark et al.[13]

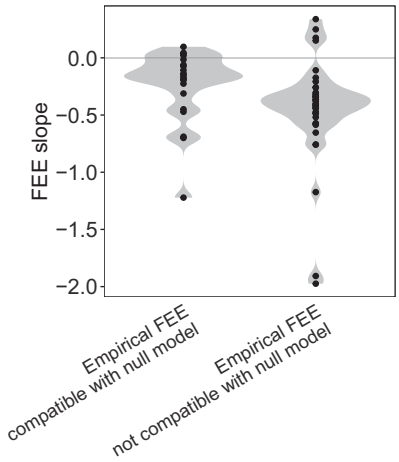


Figure S9. FEE slopes depending on compatibility with a null model, related to Figures 1 and 2

FEEs compatible with the null model (left) tend to have smaller slopes than those not compatible with the null model (right). See Methods S1 for details on the formulation of the null model.

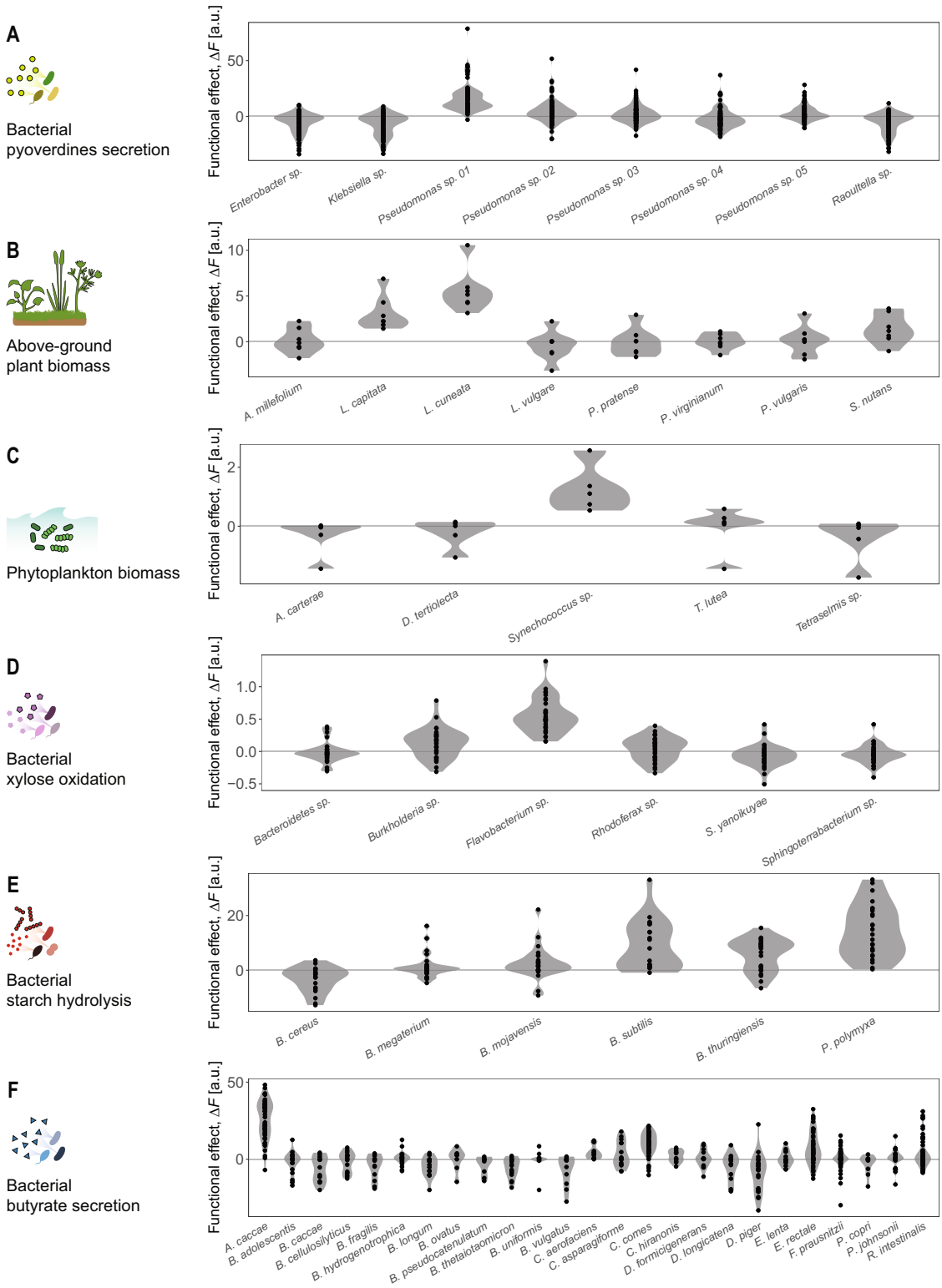


Figure S10. Distributions of species' functional effects, related to Figure 3

We quantify the functional effect of species i as $\Delta F_i(\mathbf{s}) = F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s})$, where $F(\mathbf{s})$ denotes the function of a community $\mathbf{s}$ which does not contain species i, and $F(\mathbf{s}+\mathbf{i})$ denotes the function when species i is included in the consortium. Here we represent the distributions of functional effects for all species across the empirical datasets we examined. Data sources:

(legend continued on next page)

(A) This study.
(B) Kuebbing et al.[53]
(C) Ghedini et al.[54]
(D) Langenheder et al.[52]
(E) Sanchez-Gorostiaga et al.[47]
(F) Clark et al.[13]

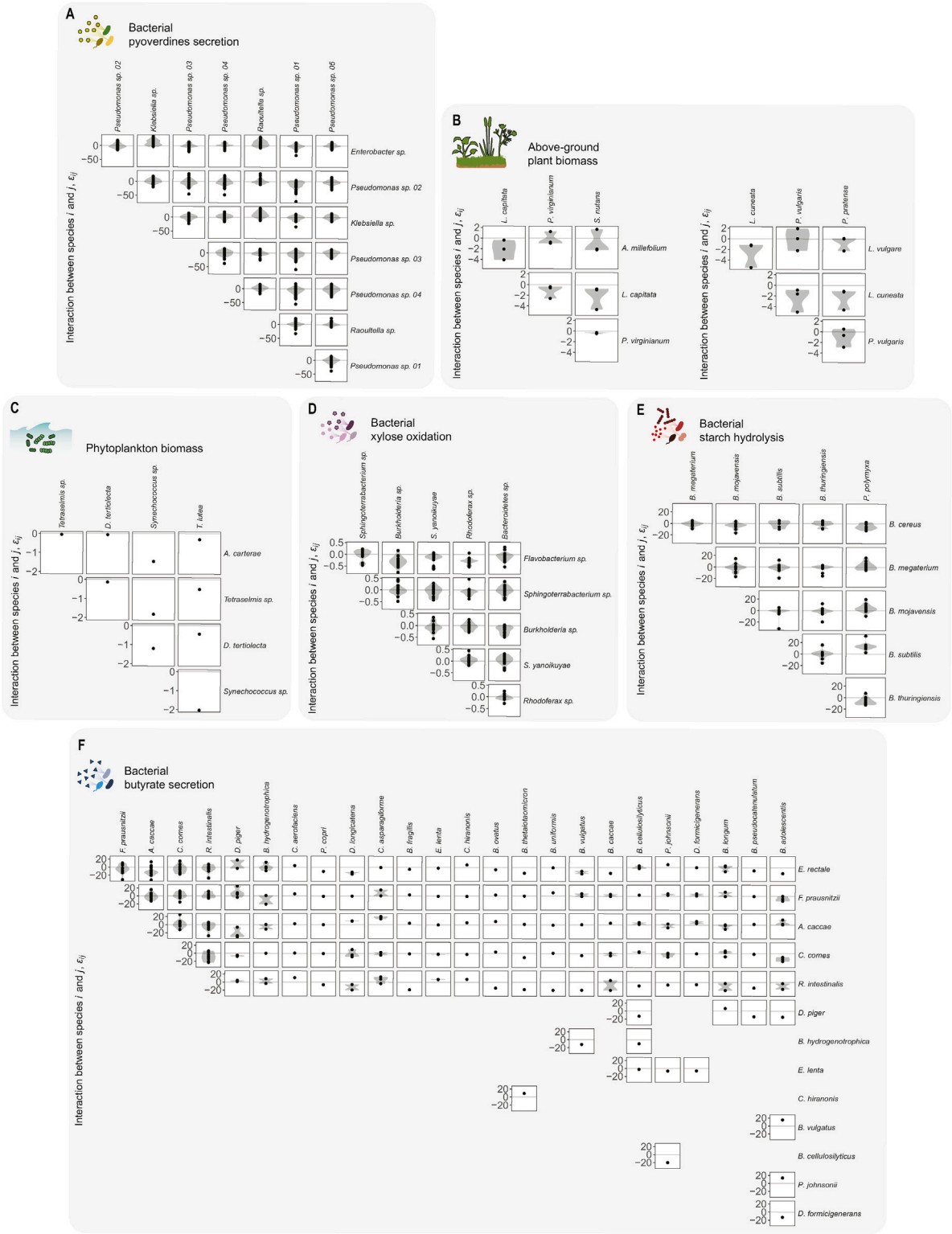


Figure S11. Distributions of epistatic effects between species, related to Figure 3

We quantify the interaction ϵij between species I and j as the deviation in function between a community containing both species with respect to the sum of their respective additive functional effects; that is: $\epsilon_{ij}(\mathbf{s}) = F(\mathbf{s}+\mathbf{i}+\mathbf{j}) - F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s}+\mathbf{j}) + F(\mathbf{s})$, where **s** is an arbitrary community not containing *i* nor *j*. Here we

(legend continued on next page)

represent the distributions of interactions (i.e., the distributions of epistatic effects) between all pairs of species across the empirical datasets we examined (note that empty panels correspond to missing data since many of the datasets are combinatorially incomplete). Data sources:
(A) This study.
(B) Kuebbing et al.[53]
(C) Ghedini et al.[54]
(D) Langenheder et al.[52]
(E) Sanchez-Gorostiaga et al.[47]
(F) Clark et al.[13]

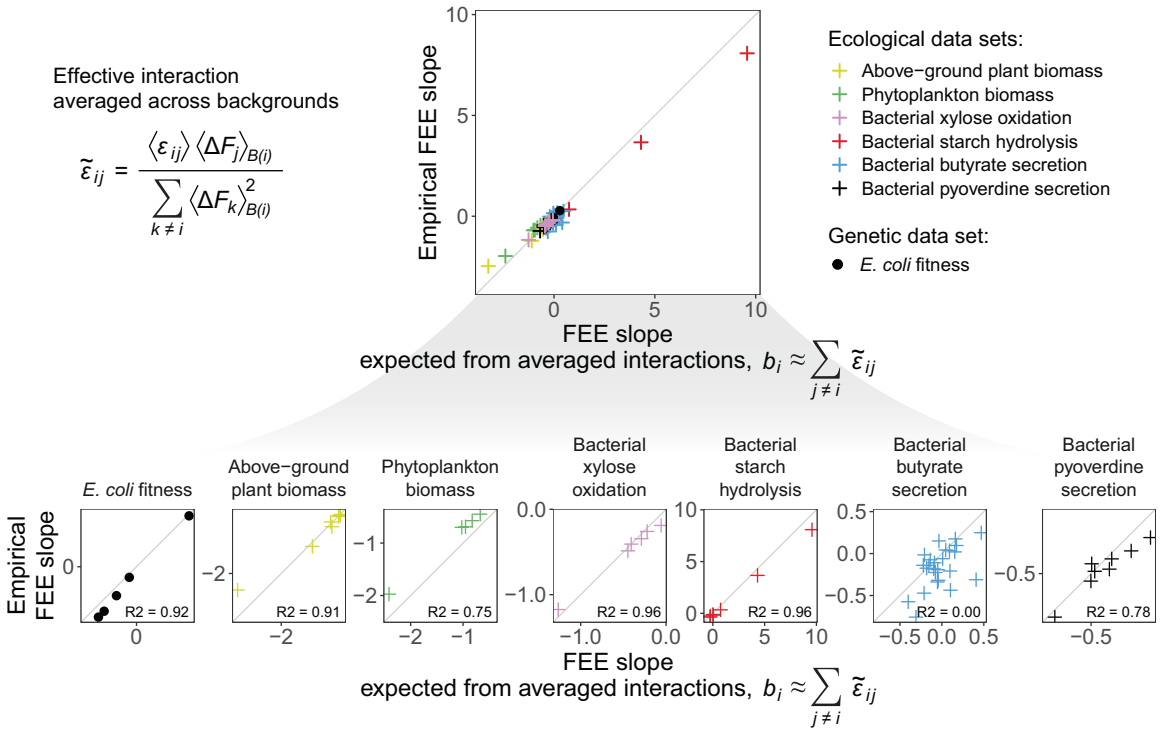


Figure S12. Effective interactions estimate FEE slopes across datasets, related to Figure 3

We define the effective interaction of species i with species j as $\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \langle \Delta F_j \rangle / \sum_{k \neq i} \langle \Delta F_k \rangle^2$ as indicated in the main text. The FEE slope for species i can be estimated as the sum of its effective interactions with all other potential community members: $b_i \approx \sum_{j \neq i} \tilde{\epsilon}_{ij}$. Here we compare the slopes estimated using this approximation with the empirically fit FEE slopes for each species, showing each dataset separately (bottom panels). The reported R^2 correspond to the $y = x$ model, that is, $R^2 = 1 - \sum_n (y_n - x_n)^2 / \sum_n (y_n - \langle y \rangle)^2$.

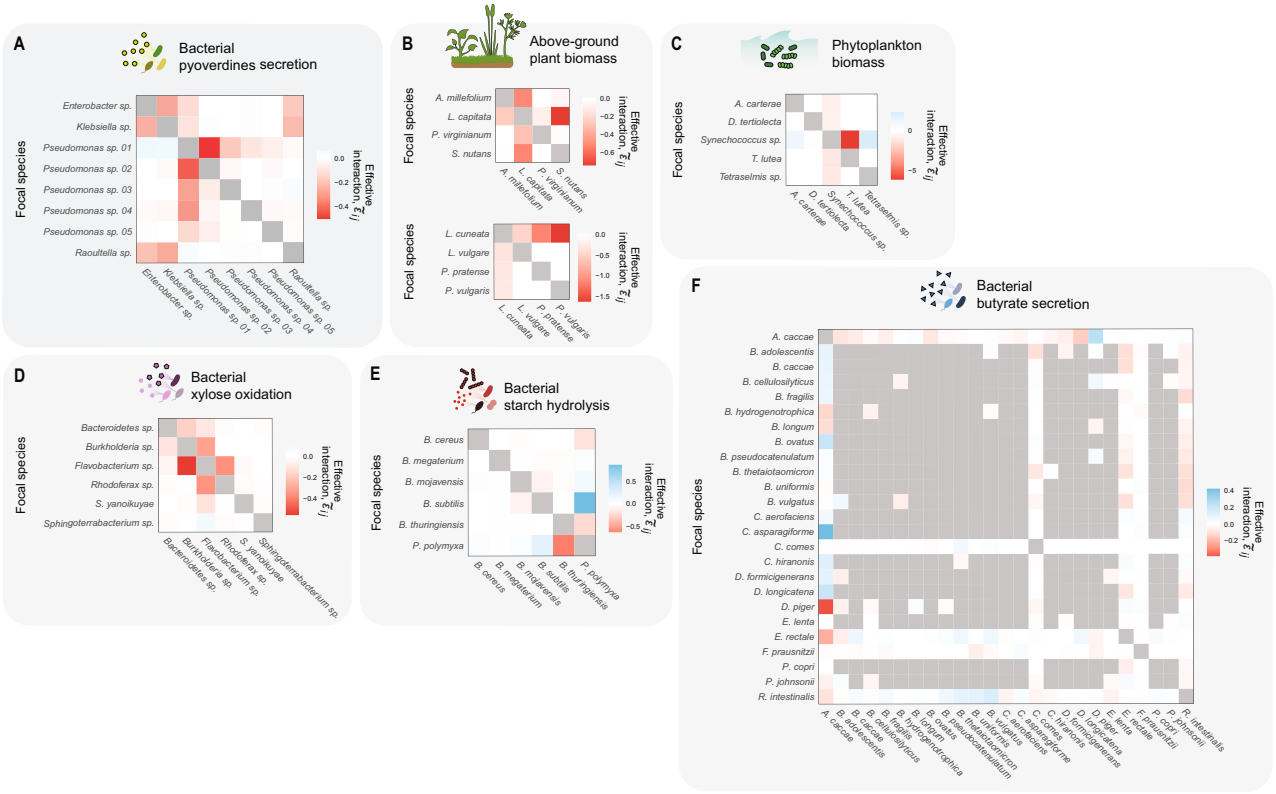


Figure S13. Effective interactions between all pairs of species, related to Figure 3

We define the effective interaction $\tilde{\epsilon}_{ij}$ of species i with species j as $\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \langle \Delta F_j \rangle / \sum_{k \neq i} \langle \Delta F_k \rangle^2$, where the averages are taken across all possible ecological backgrounds where both species i and j are not present (see main text and STAR Methods). Here we represent the magnitude of $\tilde{\epsilon}_{ij}$ for every pair of species i (focal) and j across all datasets we examined. Gray indicates that the dataset was not sufficiently complete to quantify the effective interaction. Data sources:

(A) This study.

(B) Kuebbing et al.[53]

(C) Ghedini et al.[54]

(D) Langenheder et al.[52]

(E) Sanchez-Gorostiaga et al.[47]

(F) Clark et al.[13]

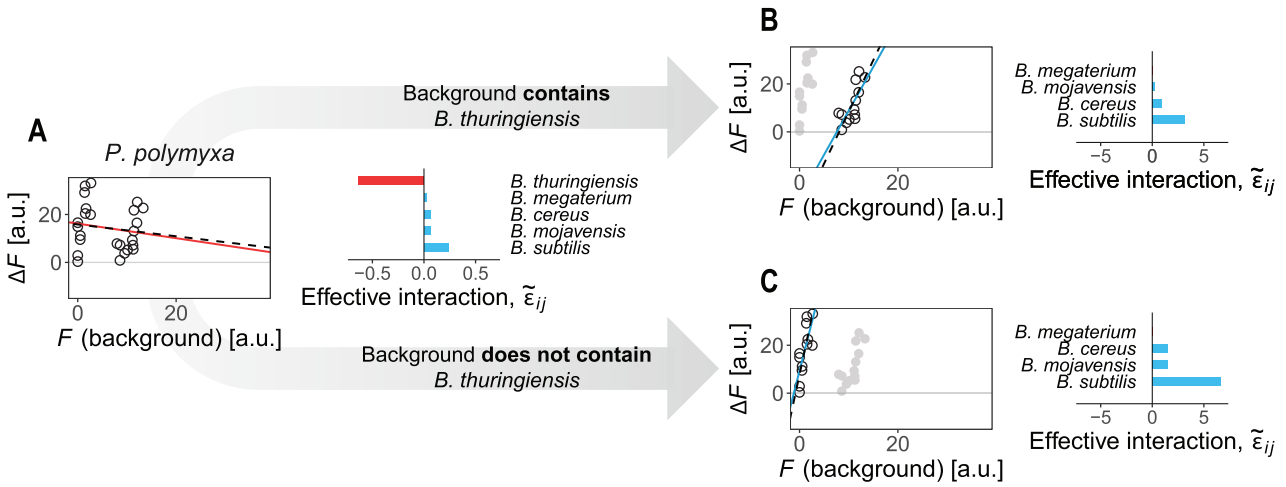


Figure S14. Effective interactions explain the branching observed in the FEE for *P. polymyxa*, related to Figure 3

(A) When we fit a single FEE for all backgrounds, we observe a negative slope. This can be explained by a strong negative effective interaction of *P. polymyxa* with *B. thuringiensis*.

(B and C) When backgrounds are split by the presence/absence of *B. thuringiensis*, the effective interactions of *P. polymyxa* with the remaining species are positive, which gives rise to the positive FEE slopes in each branch. Dashed lines represent FEEs estimated using Equation 2 and Equation S4 (see main text and Methods S1), solid lines are empirical linear fits to the data.

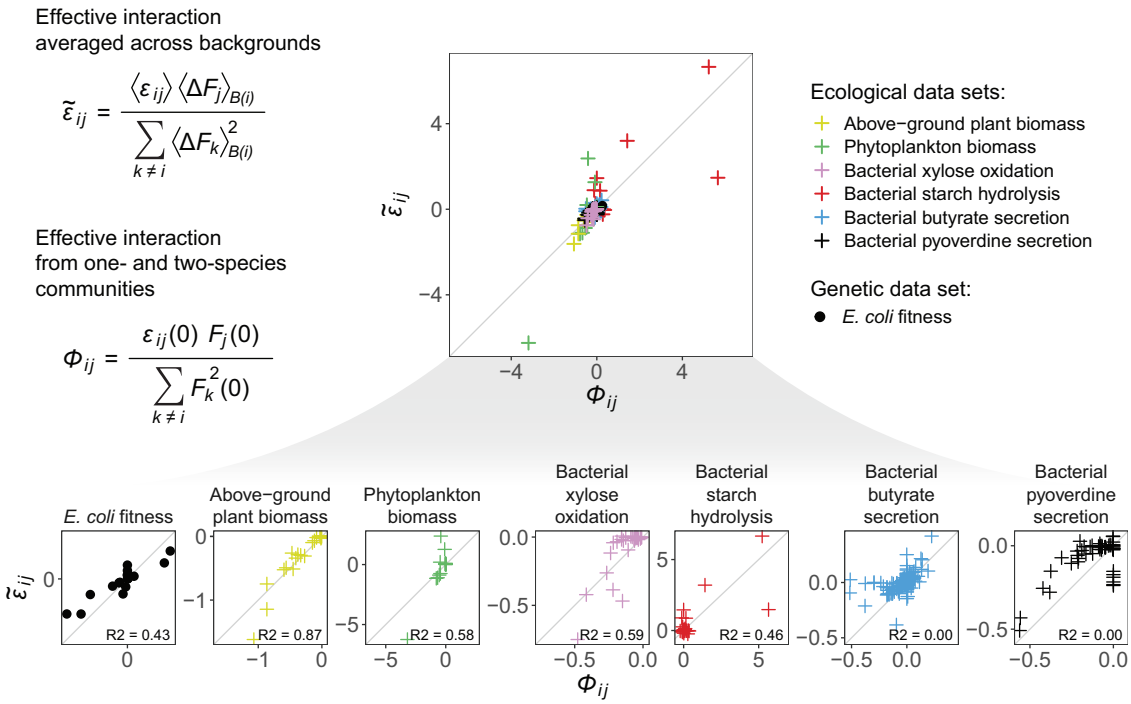


Figure S15. Effective interactions estimated from one- and two-species communities versus averaging across ecological backgrounds, related to Figure 3

We define the effective interaction of species i with species j as $\tilde{\epsilon}_{ij} \equiv \langle \epsilon_{ij} \rangle \langle \Delta F_j \rangle / \sum_{k \neq i} \langle \Delta F_k \rangle^2$ as indicated in the main text. We alternatively define $\varphi_{ij} \equiv \epsilon_{ij}(0) F_j(0) / \sum_{k \neq i} F_k^2(0)$, where $\Delta F_j(0)$ denotes the function of the monoculture of species j and $\epsilon_{ij}(0)$ denotes the deviation between the co-culture of species i and j (in the absence of additional community members) with respect to the sum of the monoculture functions (see Methods S1). The φ_{ij} are estimated just from the functions of the monocultures and two-species communities, whereas $\tilde{\epsilon}_{ij}$ are estimated by averaging interactions and functional effects across ecological backgrounds. In this figure, we compare the degree to which φ_{ij} matches $\tilde{\epsilon}_{ij}$ for every interaction where both could be quantified across datasets. The reported R^2 correspond to the $y = x$ model, that is, $R^2 = 1 - \sum_n (y_n - x_n)^2 / \sum_n (y_n - \langle y \rangle)^2$. A high R^2 indicates small higher-order effects (as is the case, for instance, for the Kuebbing et al. dataset for plant biomass[53]).

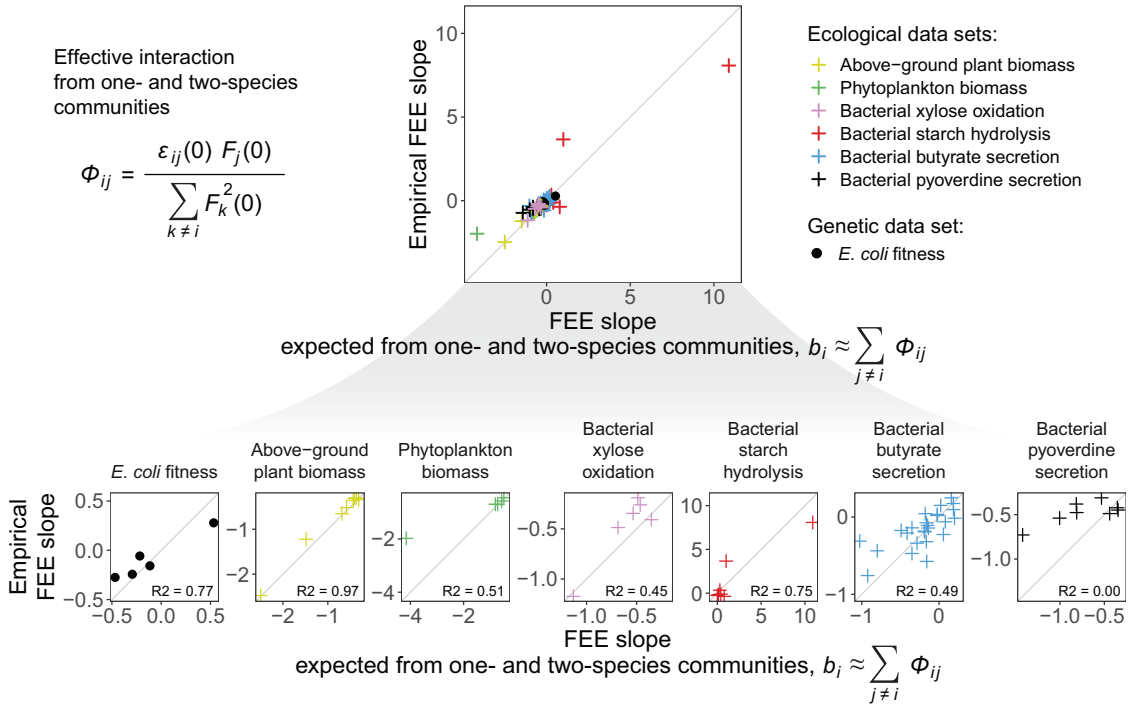


Figure S16. Estimating FEE slopes from one- and two-species communities, related to Figure 3

We define φ_{ij} as $\varphi_{ij} \equiv \epsilon_{ij}(0)F_j(0)\,/\sum_{k \neq i} F_k^2(0)$, where $F_j(0)$ denotes the function of the monoculture of species j and $\epsilon_{ij}(0)$ denotes the deviation between the co-culture of species i and j (in the absence of additional community members) with respect to the sum of the monoculture functions (see Methods S1). Here we quantify the extent to which FEE slopes (b_i for species i) can be estimated as $b_i \approx \sum_{j \neq i} \varphi_{ij}$. The reported R^2 correspond to the $y = x$ model, that is, $R^2 = 1 - \sum_n (y_n - x_n)^2 / \sum_n (y_n - \langle y \rangle)^2$. A good agreement between estimated and empirically fit FEE slopes indicates a minor role of higher-order interactions or, alternatively, that higher-order effects "cancel out" when the sum of all φ_{ij} is carried over all species j.

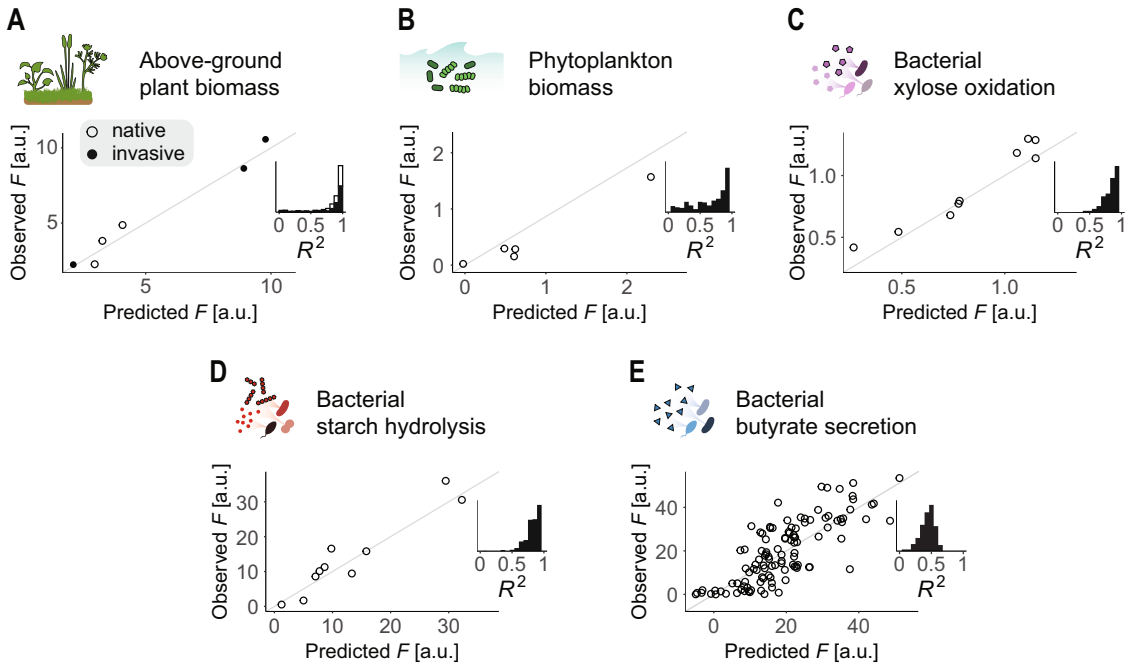


Figure S17. Predicting community function across datasets, related to Figure 4

We evaluated the ability of our statistical method (Figure 4A, STAR Methods) to predict community functions in all datasets (summarized in Table S1). For that, we left 20% of the communities in the datasets out of the sample, we used the remaining 80% to fit FEEs, and we applied our method to predict the function of the out-of-sample consortia. We quantified the accuracy of the method as the R^2 between the predictions and the observations. We repeated the same process 500 times, each leaving a different subset of communities out of sample (randomly chosen). Main plots show an example of predicted against observed functions for one of the runs. Insets show histograms of the R^2 between predictions and observations across the 500 runs.

(A) Data from Kuebbing et al.[53] Hollow/filled dots and bars correspond to native/invasive plants (note that this dataset is divided into two subsets: one of four invasive plants and another of four native plants, see Table S1).

(B) Data from Ghedini et al.[54]

(C) Data from Langenheder et al.[52]

(D) Data from Sanchez-Gorostiaga et al.[47]

(E) Data from Clark et al.[13]

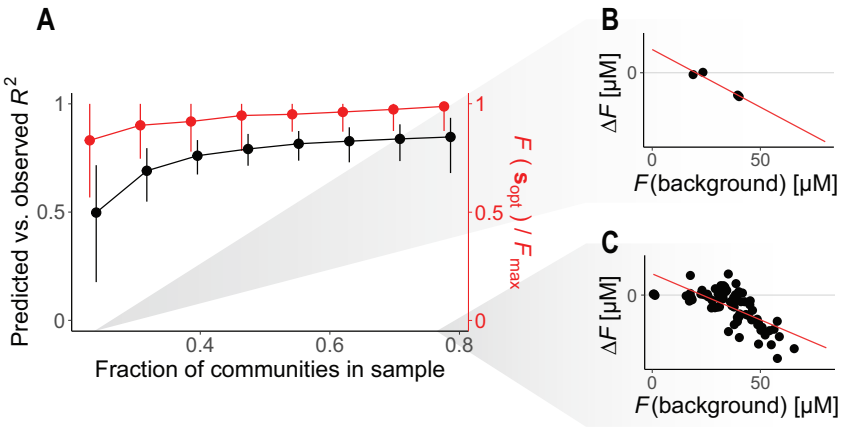


Figure S18. Predictions rely on the ability to accurately characterize FEEs, related to Figure 4

We considered all the communities in our pyoverdine experiment. We left a subset of the communities out of the sample; the remaining (in-sample) communities were used to fit FEEs, and our predictive statistical method (STAR Methods, Figure 4A) was used to predict the function of the out-of-sample communities. The quality of the predictions was quantified as the R^2 between the predicted and observed functions of the out-of-sample communities (black line). We also evaluated the empirically measured function of the assemblage predicted to be optimal (here denoted as $\mathbf{s}_{opt}$), that is, to have the highest function out of those that we tested empirically (note that there are only 225 communities tested in our dataset—164 in our first experiment and 61 additional ones in our second experiment—out of 255 total possible assemblages). We compared the measured function of this assemblage, $F(\mathbf{s}_{opt})$, with the true functional maximum, F_{max} (red line).

(A) The prediction method declines in accuracy as fewer communities are left in-sample, however, even for small sample sizes the signal remains strong and the predicted optimal community is either the true functional maximum or functionally close to it. Dots represent means, and error bars represent 95% confidence intervals across 500 runs.

(B) For the smallest sample sizes, the FEEs had to be estimated from a very small number of observations ($N \sim 4$ data points).

(C) For larger sample sizes, FEEs were estimated from a high number of observations.

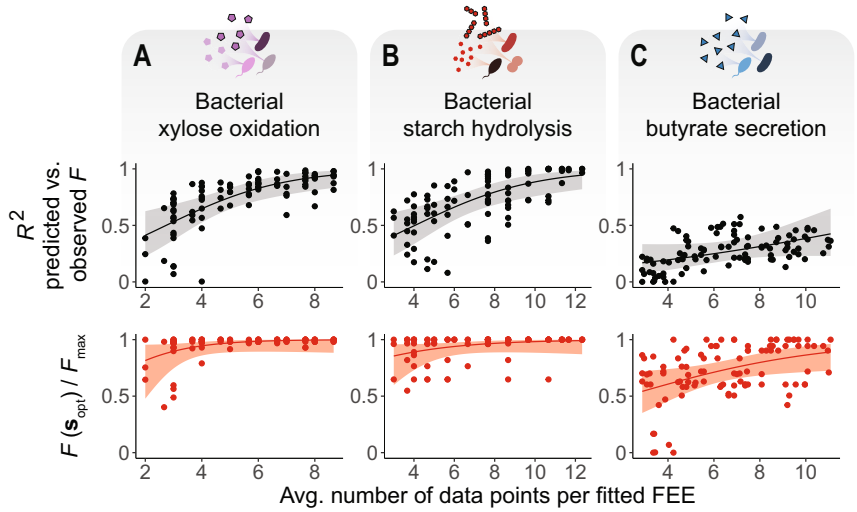


Figure S19. Prediction accuracy across datasets and sample sizes, related to Figure 4

We repeated the analysis in Figure S18 for the three other datasets with the largest combinatorial size. We represent the number of data points used to fit each FEE, averaged across all species, against the R^2 between the predicted and measured functions of the out-of-sample communities (black), and against the function of the predicted functional maximum ($F(\mathbf{s}_{opt})$) relative to the true maximum function (F_{max}) across all measured assemblages (red).

(A) Data from Langenheder et al.[52]

(B) Data from Sanchez-Gorostiaga et al.[47]

(C) Data from Clark et al.[13]

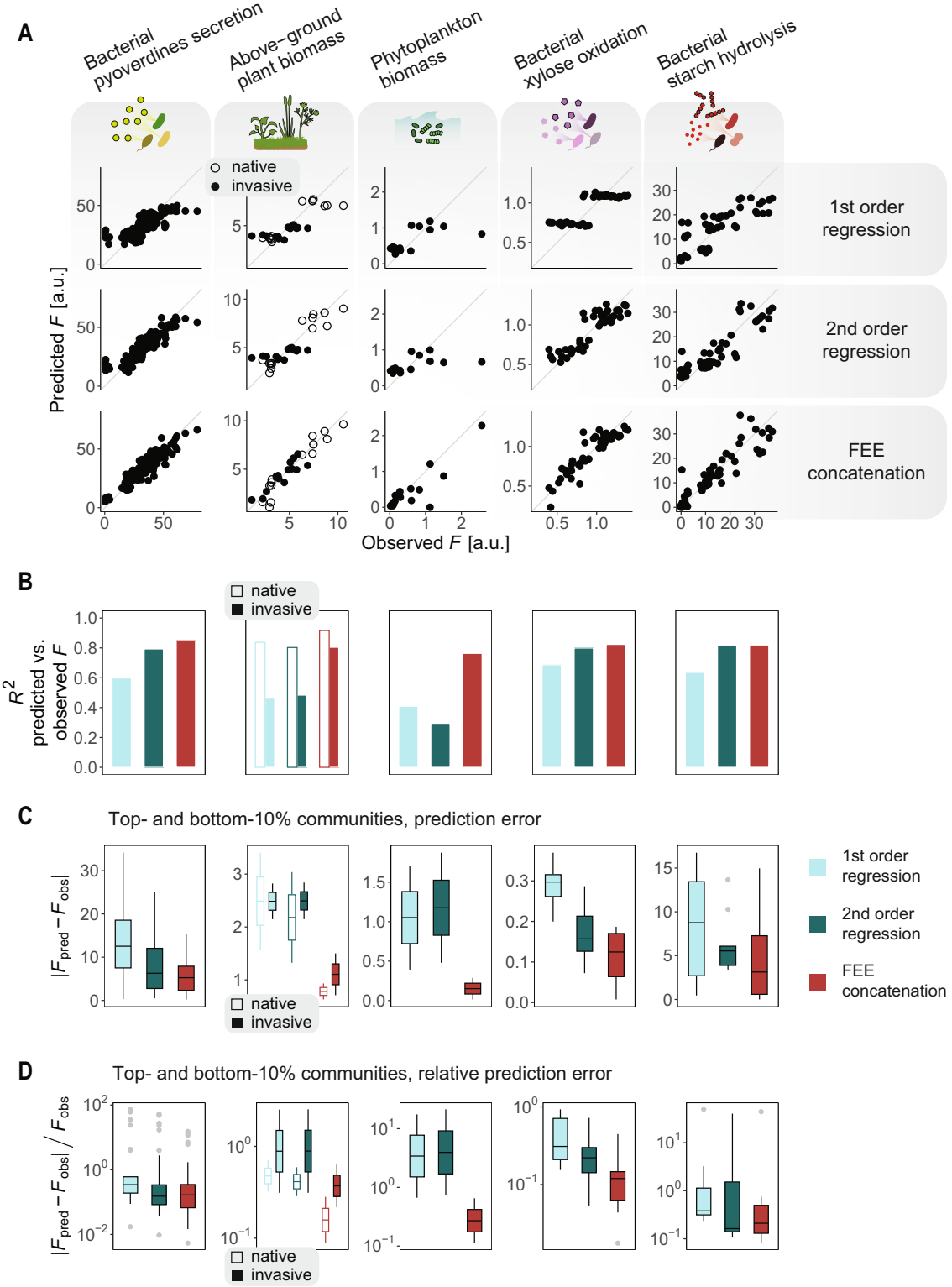


Figure S20. Predictions of community function based on FEE concatenation are more accurate than alternative methods, related to Figure 4
We performed leave-one-out cross-validations across various datasets to compare the performance of our predictive method with respect to first- and second-order regression models.

(legend continued on next page)

(A) Each column of the grid corresponds to a different dataset, and each row corresponds to a different method for predicting out-of-sample community functions. Note that in the Kuebbing et al. dataset[53] there are two subsets of species (a set of 4 native plants and another set of 4 invasive plants, see Table S1), which were treated as separate datasets despite being represented together (with hollow/filled dots and bars, respectively).
(B) The performance of each method was quantified as the R^2 between the predicted and observed values for the functions. Our method consistently performed better than, or as good as, first- and second-order regressions.
(C and D) Absolute and relative prediction error of each model (blue: first- and second-order regression, red: FEE concatenation) within the set of communities with the highest or lowest functions (top and bottom 10%).

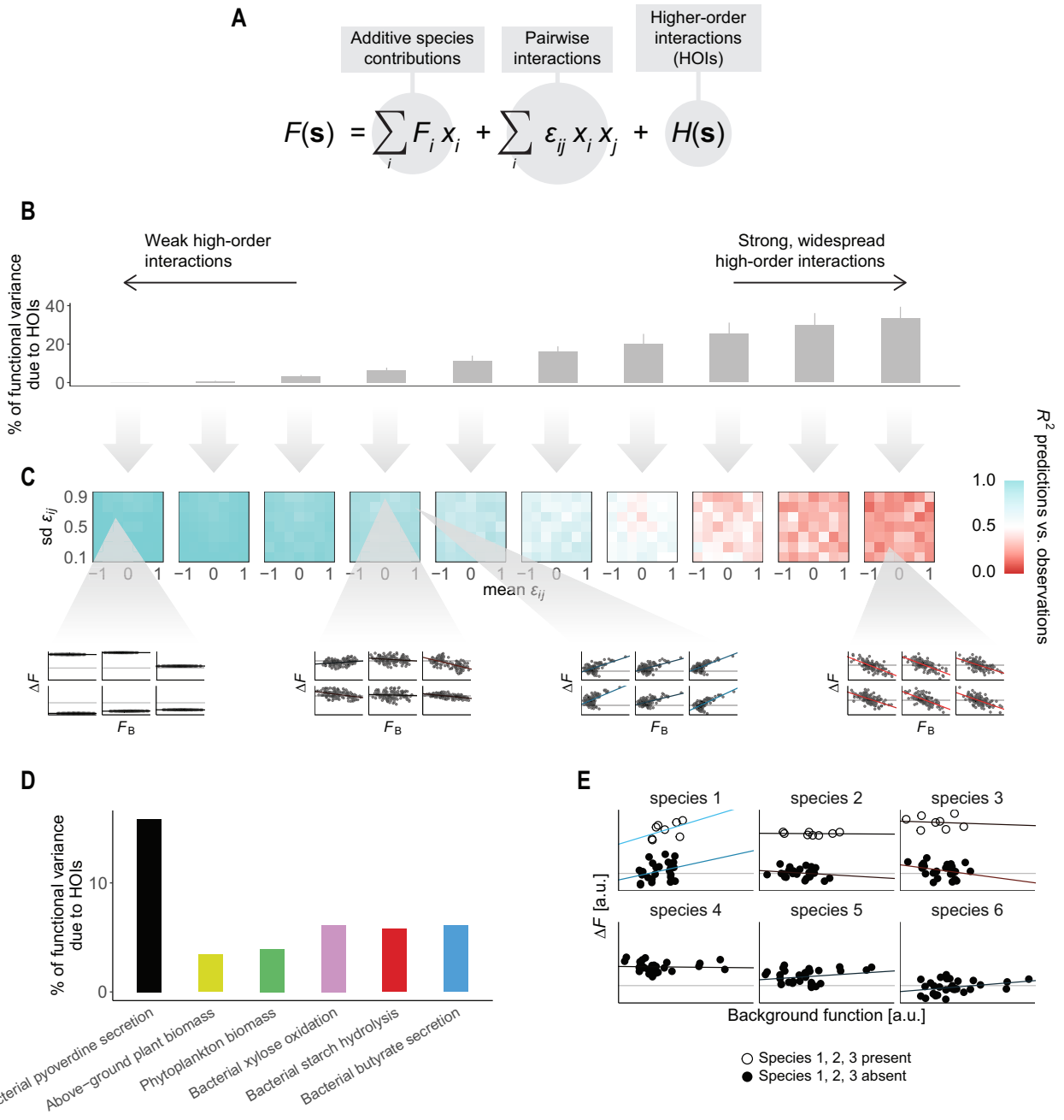


Figure S21. Performance of FEE-based prediction method on simulated data, related to Figure 4

We simulated a set of mappings between community composition and function where ecological interactions were of variable sign, magnitude, and order.

(A) Formulation of the model from which community structure-function mappings were simulated.

(B) Modulating the magnitude of the terms $H(\mathbf{s})$ allows us to vary the fraction of variance across community functions, which can be attributed to higher-order interactions (HOIs).

(C) We quantify the performance of our FEE-based method as the R^2 between the true and predicted functional values (via leave-one-out cross-validation).

(D) We show the estimated percentage of functional variance due to higher-order interactions in the empirical datasets we examined in this study. This percentage was computed by fitting a second-order regression model to each full dataset and quantifying the fraction of variance unexplained by the model.

(E) Example of FEEs emerging in a simulated dataset where we consider a single, strong high-order interaction between species 1, 2, and 3. Note that FEEs can still predict the functional effect of each species, but the FEE for species 1, 2, and 3 appears "split" into two branches—each branch is defined by the presence (hollow dots) or absence (filled dots) of the three species involved in the HOI.

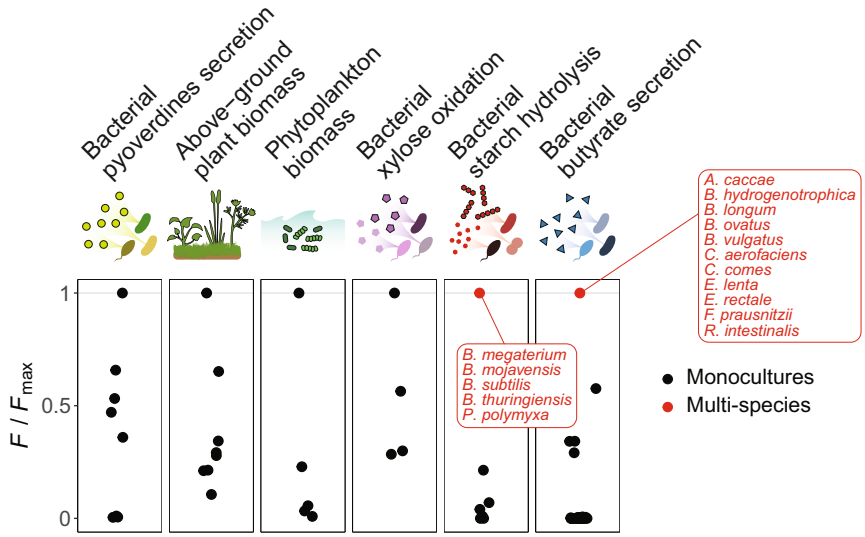


Figure S22. Optimal communities across datasets, related to Figures 1 and 2

For each dataset, we represent the functions of the monocultures (black dots) with respect to the maximum function observed across all consortia (F_{max}). Whenever the functional maximum corresponds to a multi-species assemblage, it is also represented in red. Note that not all datasets are combinatorially complete, so the possibility that a multi-species community is the true functional maximum (which was not empirically tested) cannot be ruled out even if the one represented here is a monoculture.